DATA INTENSIVE COMPUTING

CSE 587

FINAL PROJECT: COMMUNICATING THE RESULTS OF DATA ANALYTICS

Saurabh Pradip Bajoria 50208005

Vidhi Jatin Shah 50207090

Data Source

We have selected **Indian Premier League** data over 9 years (2008-2016) from Kaggle.
It has the following two files:

1. **matches.csv:** It contains metadata for each match such as the Teams participating, Venue, Toss winner, Match winner etc.
2. **deliveries.csv:** It contains ball-by-ball data for every match played.

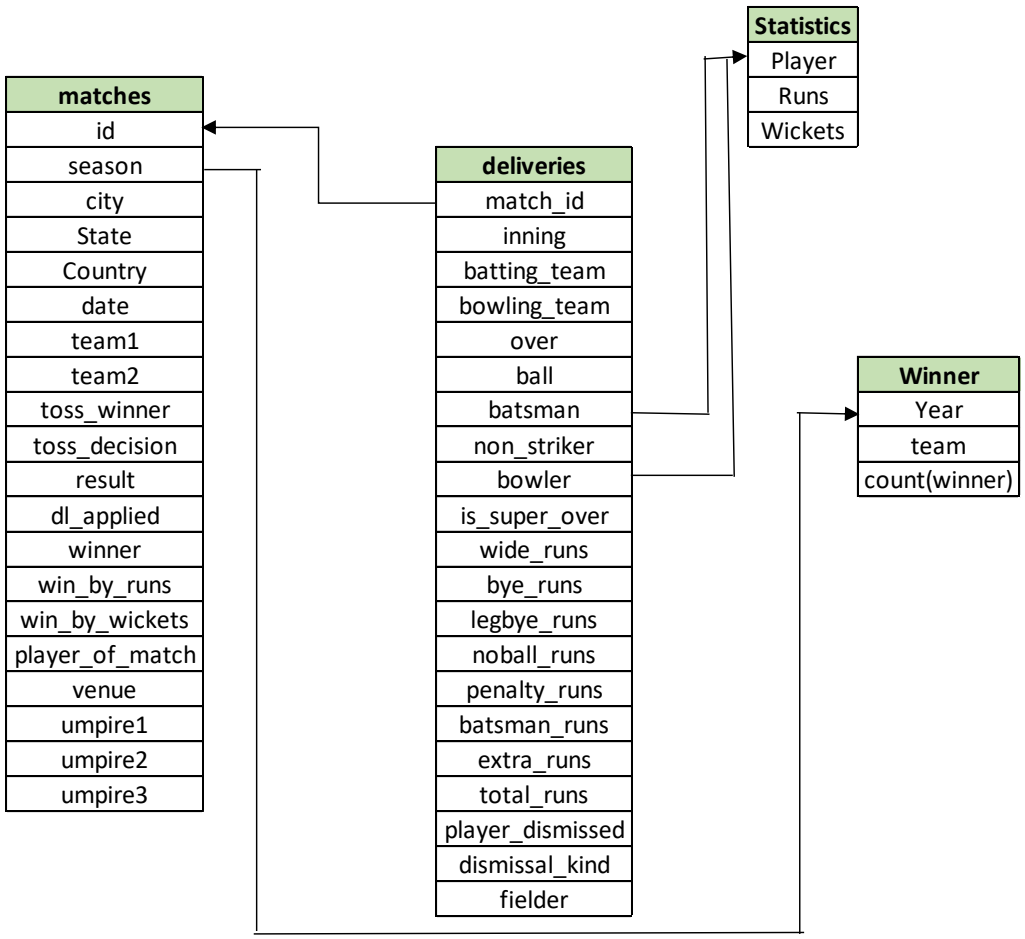
- These primary key matches.Id has a one to many relation with the deliveries.MatchId and corresponds to every match played across all the seasons of the IPL.

We have also created the following two csv files using the above data and R-code.

1. **Statistics.csv:** It contains the total runs scored and total wickets taken by every player.
2. **Winner.csv:** It contains the number of wins per team per year.

- Please find the Rcode for the generation of the files along with the submission.

Data Schema:

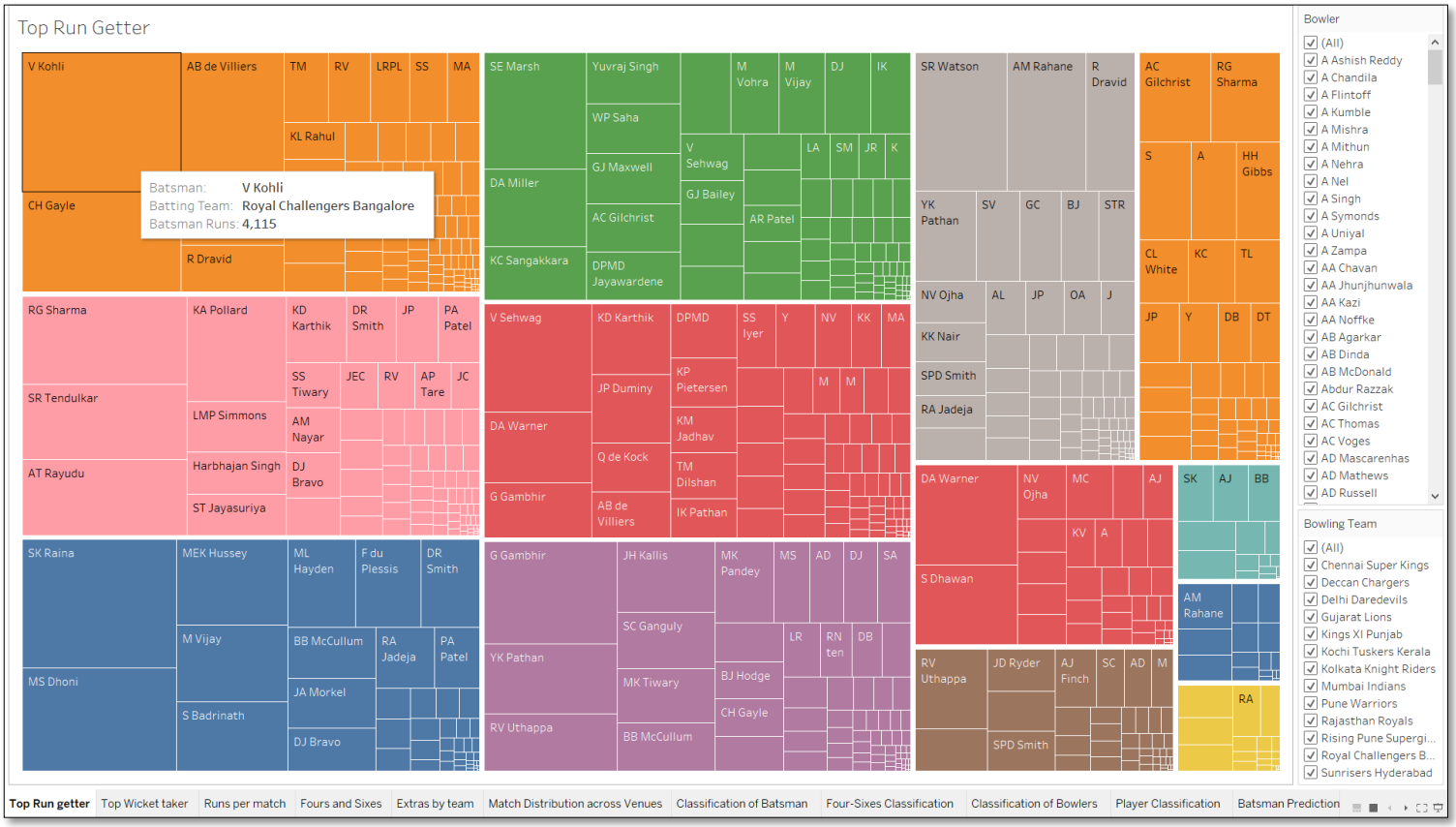


Activity 2: Story telling with your data

We have created the following Worksheets and Dashboards:

Worksheets:

1. Top Run Getter:



Plot (1)

The above Treemaps plot (1) show the top run scorers for each team across all seasons of the IPL (Indian Premier League).

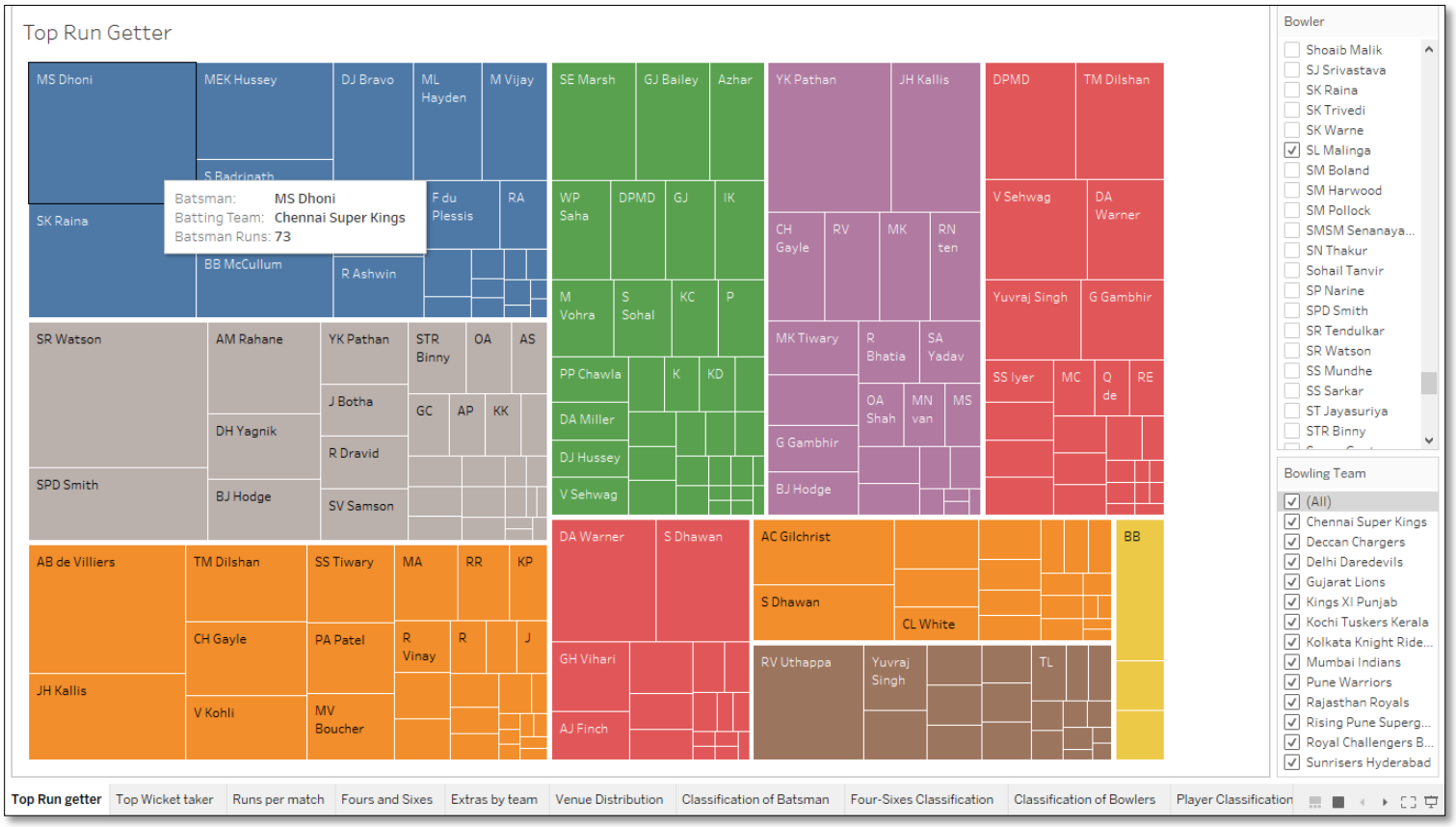
Fields Used:

1. **Batsman Runs (deliveries.csv):** Aggregated using Batsman Runs
2. **Batting Team (deliveries.csv):** Colored all the players in a team using this field.
3. **Batsman (deliveries.csv):** Provided the Labels to top players using this field.

Filters:

1. **Bowler (deliveries.csv):** Provided a filter to change the bowler (bowlers).
   If a bowler is selected, the plots shows the runs scored by each batsman against that bowler.
   For e.g. in the below plot (2), Bowler is filtered to **SL Malinga.** The plot shows the runs scored by each batsman against SL Malinga. **MS Dhoni** tops the list with 73 runs scored for **Chennai Super Kings** team, as seen in the tool tip.

2. **Bowling Team (deliveries.csv):** This filter gives the total number of runs scored by each player against a bowling team.



Plot (2)

2. Top Wicket taker:

The below **Horizontal Bars** plot (3) show the top wicket takers across all seasons of the IPL (Indian Premier League).

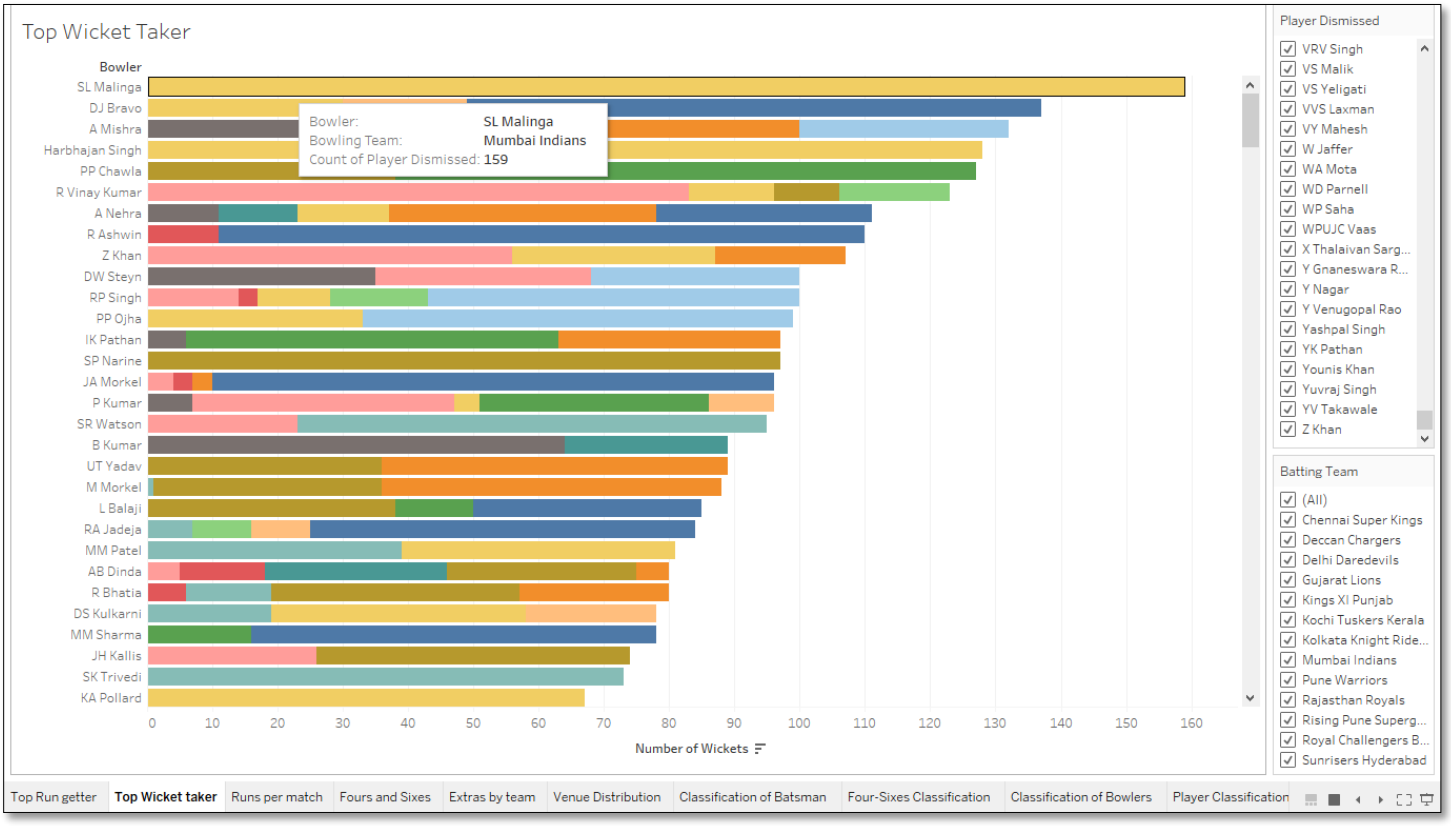
Fields Used:

1. **Player Dismissed (deliveries.csv):** Counted the number of player dismissed to get the total number of wickets taken by each bowler.
2. **Bowling Team (deliveries.csv):** Colored all the players in a team using this field.
3. **Bowler (deliveries.csv):** This gives a list of bowlers.

Filters:

1. **Batting Team (deliveries.csv):** This filter gives the total number of wickets taken by each player against a batting team.

2. **Dismissal Kind (deliveries.csv):** This filters the wickets taken by a bowler of a specific dismissal type i.e. bowled, catch out etc.

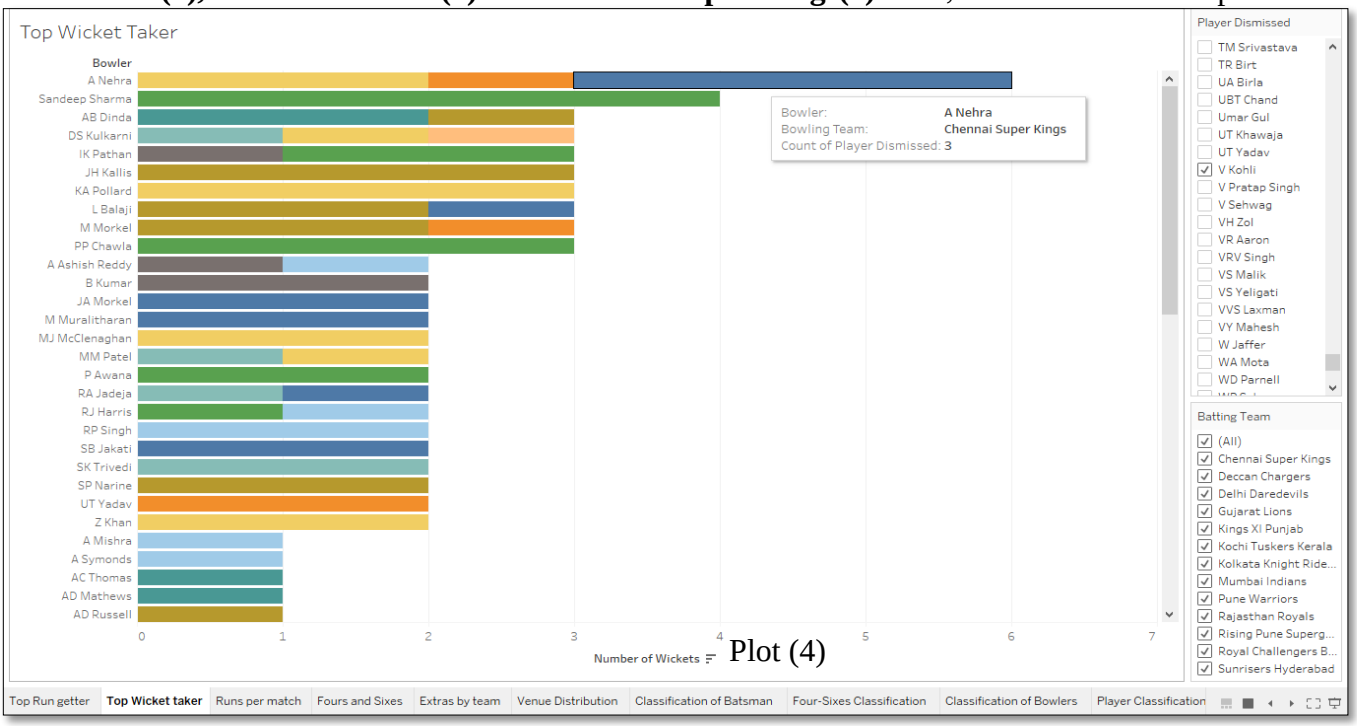


Plot (3)

3. **Player Dismissed (deliveries.csv):** Provided a filter to change the Batman (batsmen).

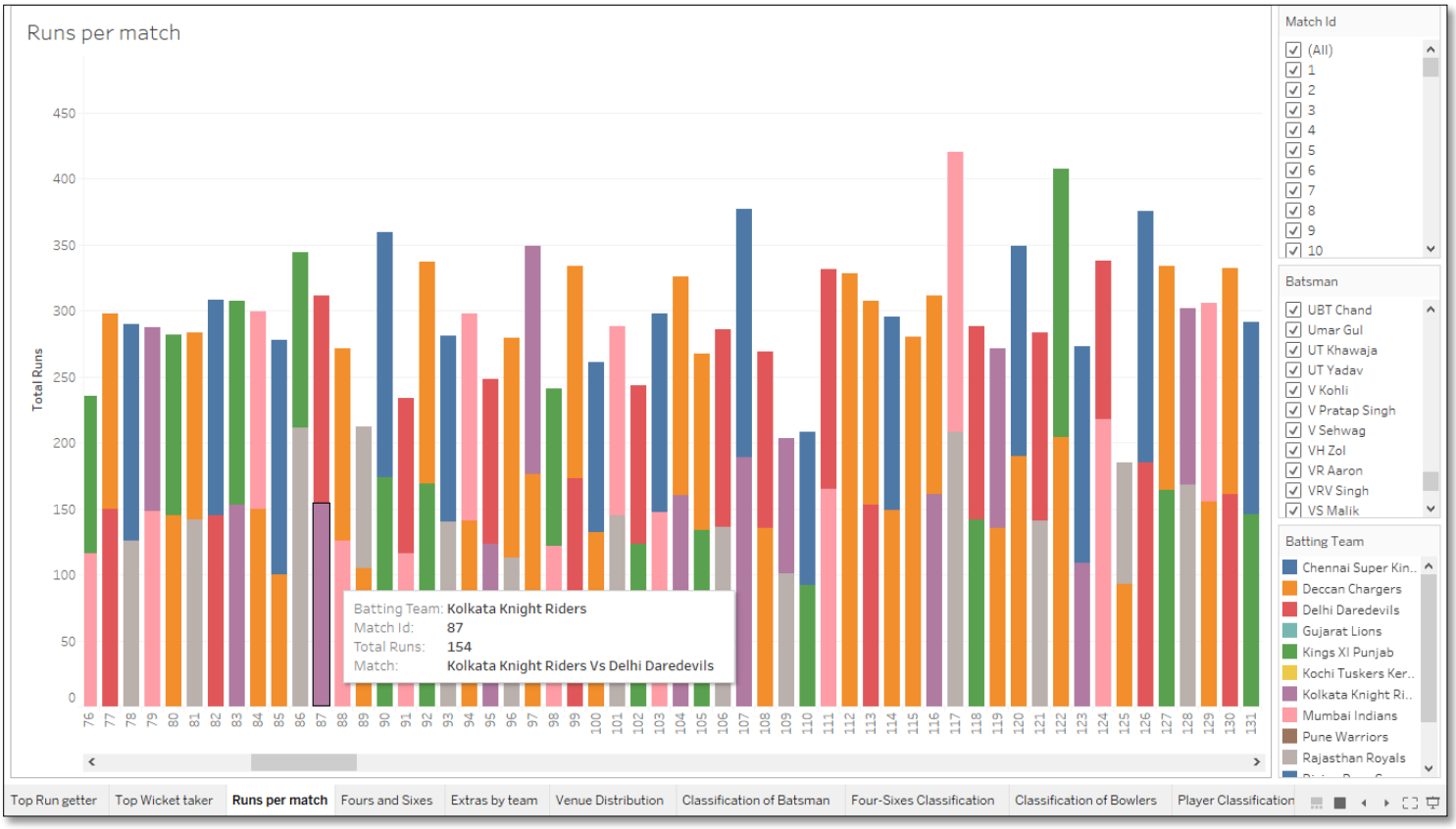
If a batsman is selected, the plot gives the number of times each bowler has dismissed that batsman.

For e.g. in the below plot (4), Player Dismissed is filtered to **V Kohli.** The plot shows the number of times each bowler has dismissed V Kohli. **A Nehra** tops the list with 6 wickets for **Mumbai Indians(2), Delhi Daredevils(1) and Chennai Super Kings(3)** team, as seen in the tool tip.



Plot (4)

3. Runs Per Match:



Plot (5)

The above **Stacked Bars** plot (5) show total runs scored per match by both the teams across all seasons of the IPL (Indian Premier League).

Fields Used:

1. **Total Runs (deliveries.csv):** Aggregated the field Total Runs to calculate the total number of runs scored by each team in a match.
2. **Batting Team (deliveries.csv):** Differentiated the runs scored by each team in a match using the field Batting Team.
3. **Match Id (matches.csv):** This gives the list of Matches played over all the seasons.

Fields Created:

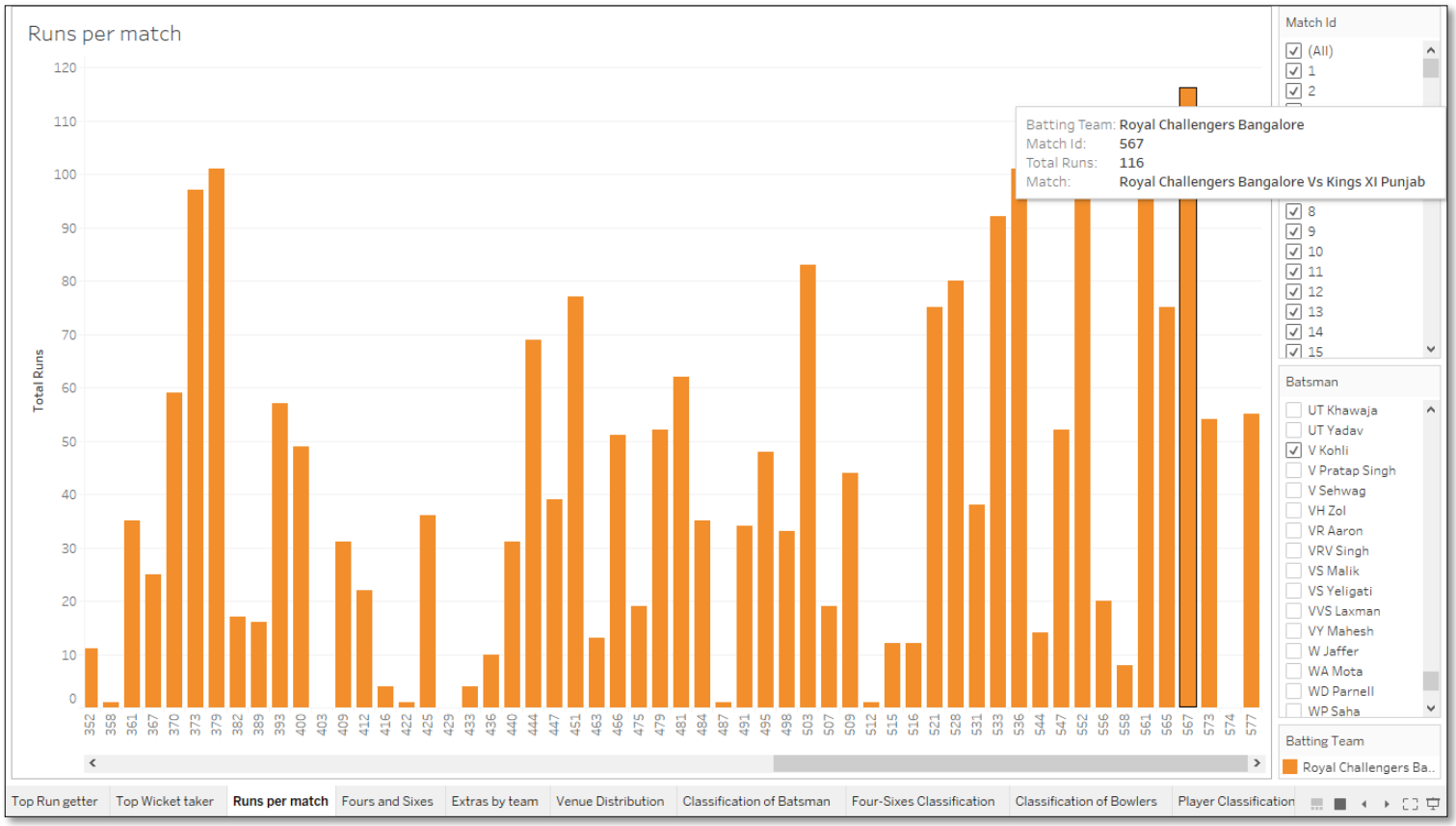
1. **Match:** This field gives the details of the teams participating in a match that corresponds to the Match Id.

   Formula: ATTR([Team1]) + " Vs " + ATTR([Team1])

Filters:

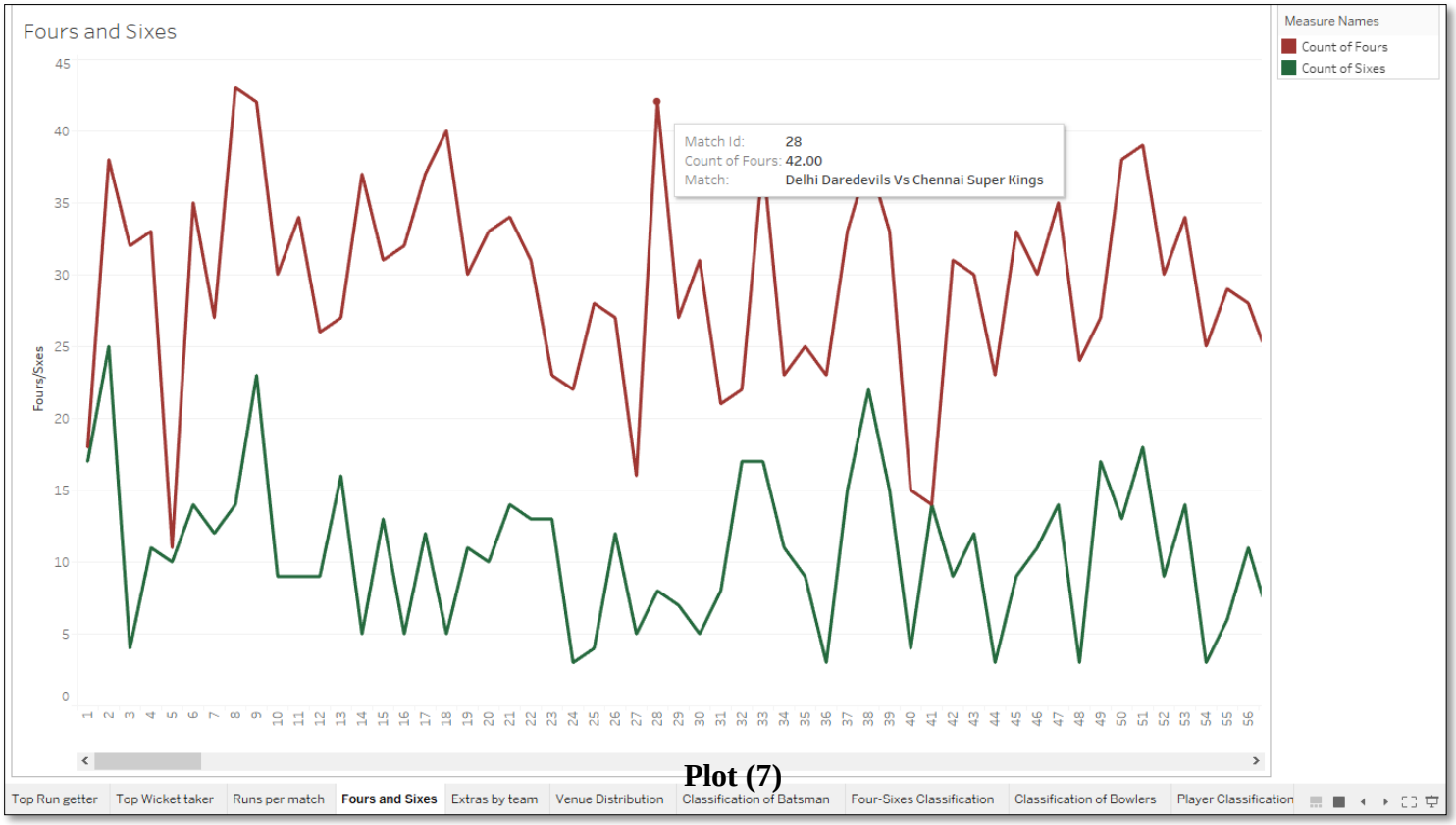
1. **Match Id (matches.csv):** Provided a filter to change the Match.

2. **Batsman (deliveries.csv**): If a batsman is selected, the plot gives the number of runs scored by that batsman across all the matches. For e.g. in the below plot (6), **V Kohli** is selected as Batsman. The plot shows the number of runs scored by V Kohli across all the matches that he has played. The color of the plot shows the team for which he played in those matches.



Plot (6)

4. Fours and Sixes:



Plot (7)

The above **Line Graph** plot (7) show the number of fours and sixes scored per match by both the teams combined across all seasons of the IPL (Indian Premier League).

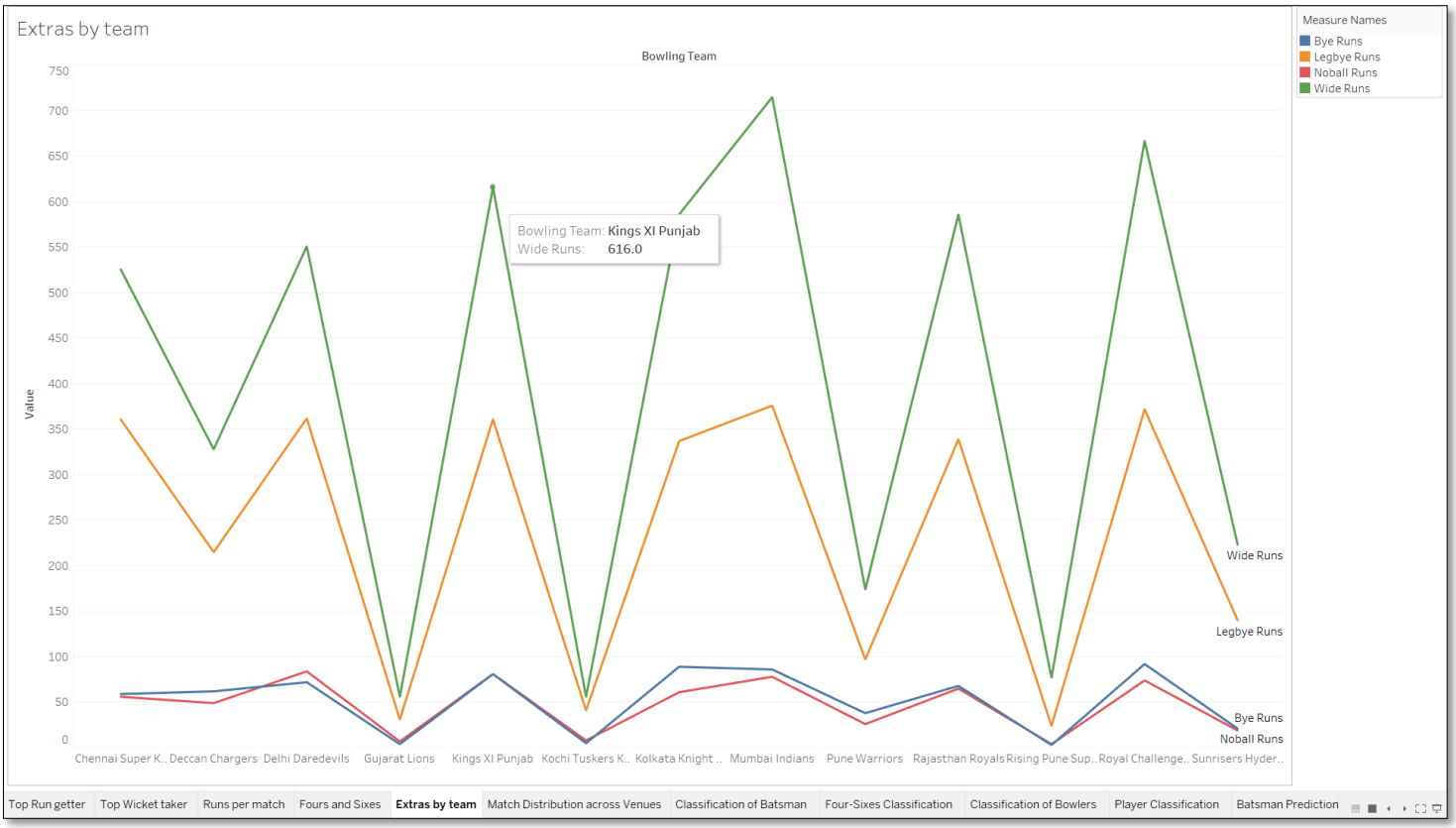
Fields Used:

1. **Match Id (matches.csv):** This gives the list of Matches played over all the seasons.

Fields Created:

1. **Fours:** This field gives the number of fours hit.
   Formula: IF [Batsman Runs] ==4 Then 1 END
2. **Sixes:** This field gives the number of sixes hit.
   Formula: IF [Batsman Runs] ==6 Then 1 END

5. Extras by team:



Plot (8)

The above **Line Graph** plot (8) show the number of extras (Wides, No Balls, Leg Byes and Byes) conceded by a team.

Fields Used:

1. **Bowling team(deliveries.csv):** This gives the list of bowling teams.
2. **Bye Runs (deliveries.csv):** Aggregated this field to get the total number of bye runs given by a team.
3. **Legbye Runs (deliveries.csv):** Aggregated this field to get the total number of Legbye runs given by a team.
4. **Wide Runs (deliveries.csv):** Aggregated this field to get the total number of Wide runs given by a team.

5. **Noball Runs (deliveries.csv):** Aggregated this field to get the total number of Noball runs given by a team.

- Each type of extras is given a different color to make it visually appealing.

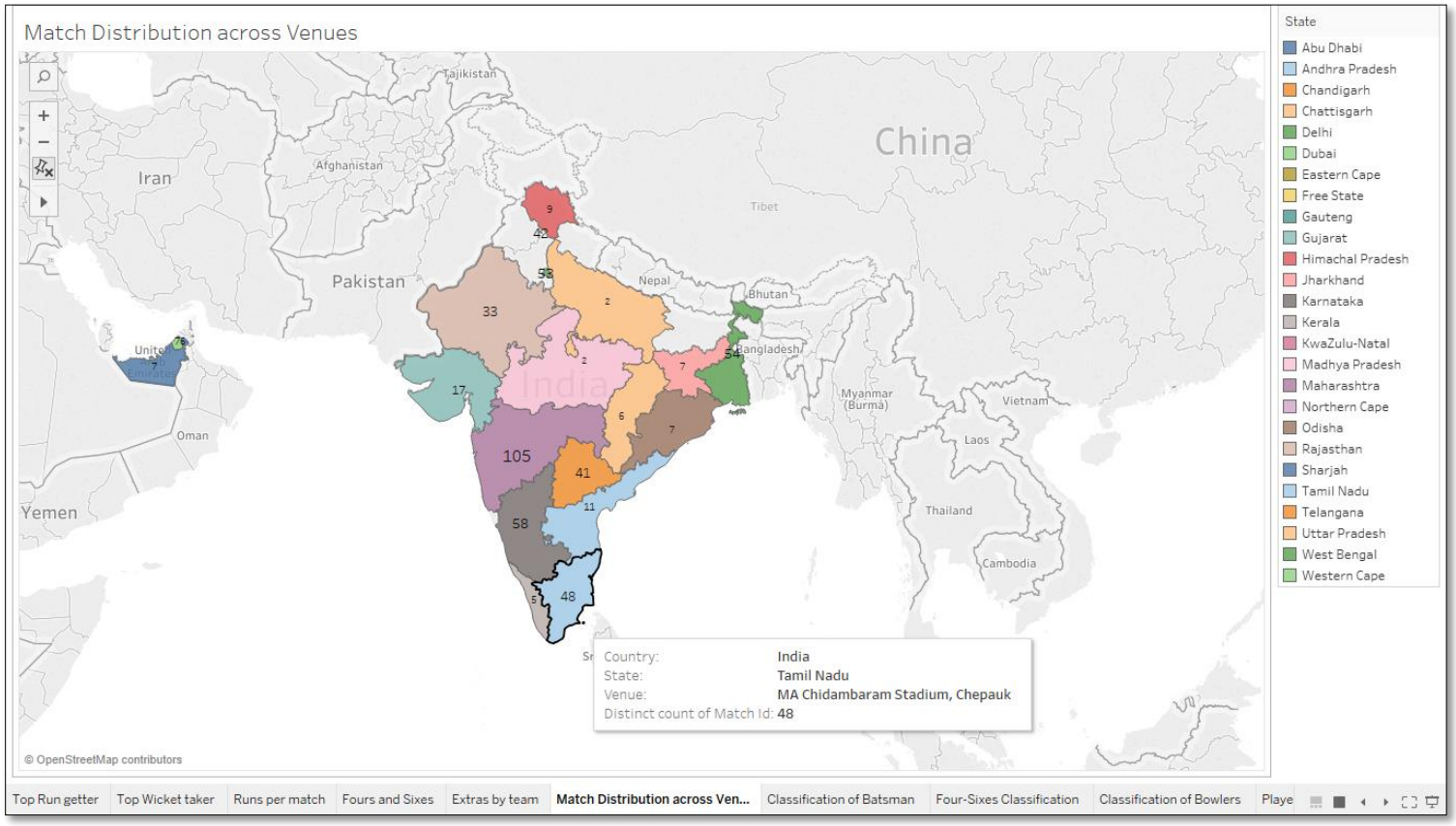
6. Match distribution across Venues:

The below **Filled Map** plot (9) show the total number of matches played at different Venues in a State.

Fields Used:

1. **Match Id(deliveries.csv):** Aggregated using the Match ID to give the total number of matches played in a State.
2. **Venue (matches.csv):** Used a tool tip to show the venue in a state.
3. **State (matches.csv):** Provides the state of the venue at which the match is played.
4. **Country (matches.csv):** Provides the Country of the venue at which the match is played.

- **Latitude and Longitude** are generated automatically using State and Country.
- Each state is given a different color and the size differs depending on the number of matches played in that state. This make the plot visually appealing.



Plot (9)

Dashboard:

Player Analysis Dashboard:

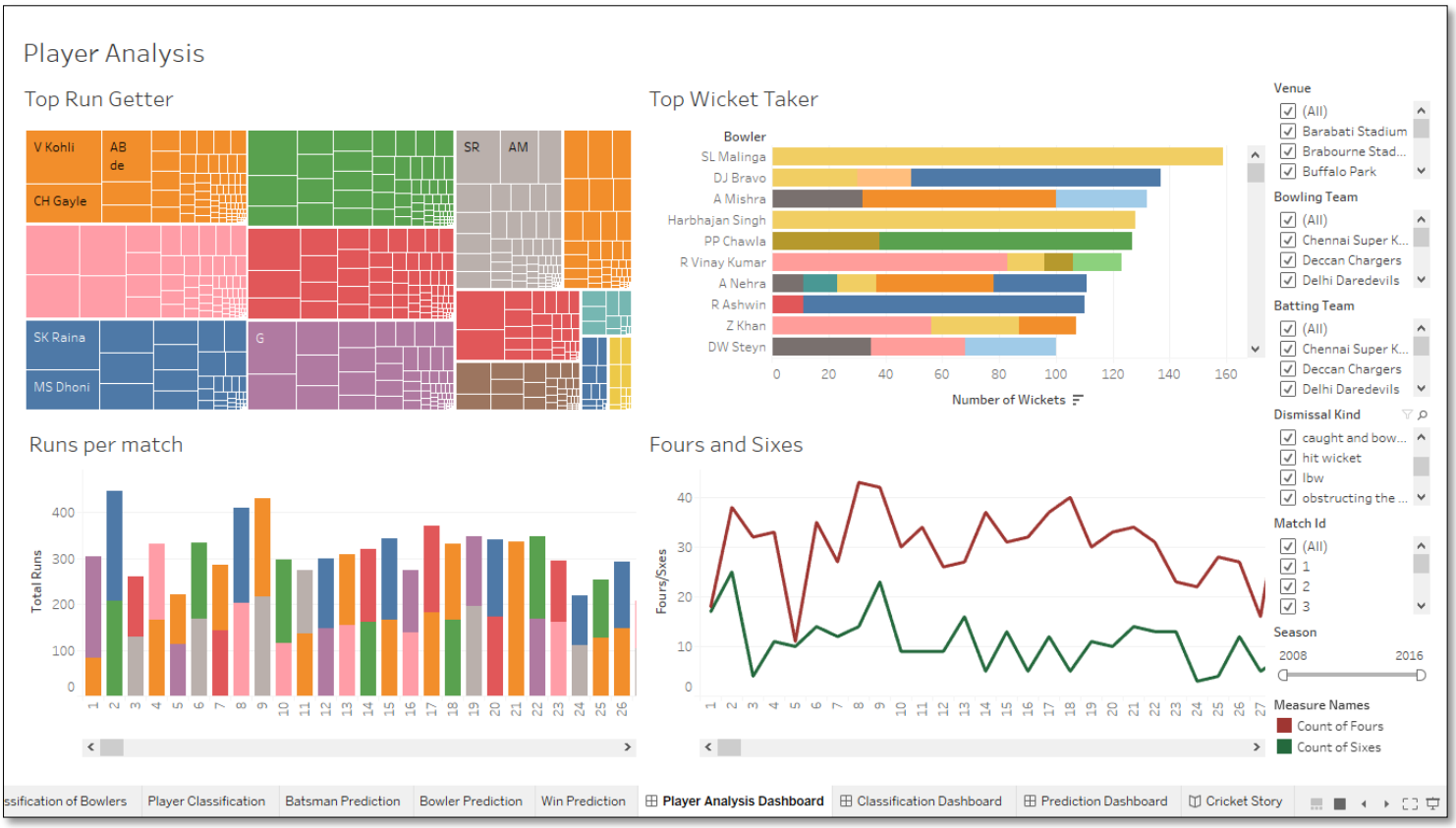
The below Dashboard plot (10) provides overall analysis of the players across all the seasons of the IPL.

Worksheets used:

1. **Top Run getter**
2. **Top Wicket Taker**
3. **Runs per match**
4. **Fours and Sixes**

Filters:

1. **Venue (matches.csv):** Provides the statistics (runs scored and wickets taken) at a venue.
2. **Bowling team (deliveries.csv):** Filters all the statistics for/against a bowling team.
3. **Batting team (deliveries.csv):** Filters all the statistics for/against a batting team.
4. **Dismissal kind (deliveries.csv):** Filters the wickets taken by a bowler pertaining to a specific kind.
5. **Match id (deliveries.csv):** Filters the records of one Match or several matches.
6. **Season (matches.csv):** Filters the records of one/more seasons.



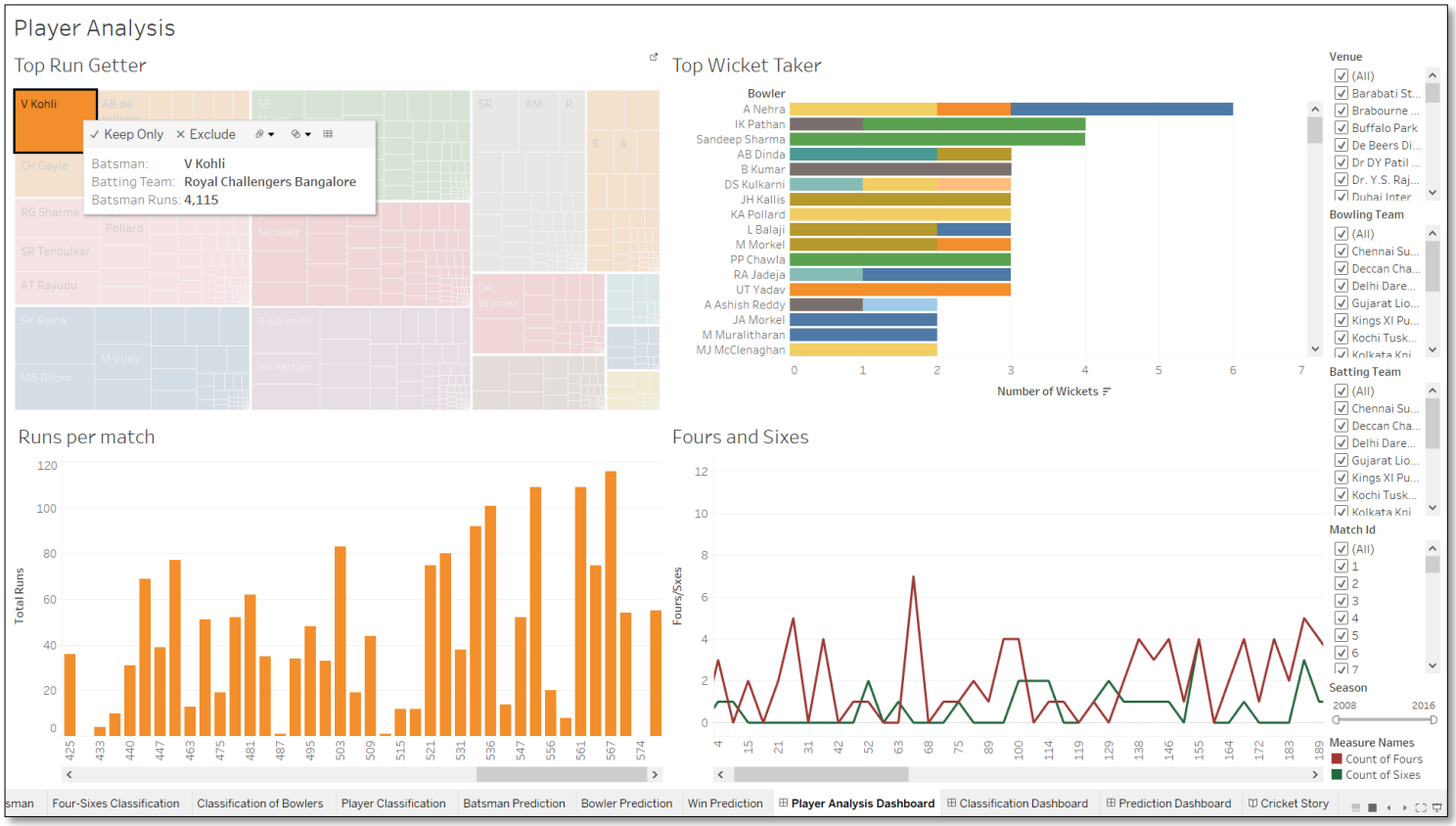
Plot (10)

Description/Usage:

Player Analysis dashboard can be used for the following purposes:

1. Analyzing a player based on the
   - ➢ total runs scored
   - ➢ total wickets taken
2. Analyzing a team based on the
   - ➢ runs scored per match
   - ➢ fours and sixes hit

We can also dig the analysis deep by using the filters provided. For e.g. we can get the runs scored, wickets taken and fours and sixes hit by a player in all the matches that he has played.
The user can filter the player using the explicit filters provided or by clicking on any player in any plot. This is illustrated below:



Plot (11)

As seen in the above dashboard Plot (11), filter is applied by clicking on the Player V Kohli in the **Top Run Getter** plot. This applies a filter of batsman across all the others plots in the dashboard.

- ➢ **Top Wicket Taker** plot now shows the number of times each bowler has dismissed V Kohli.
- ➢ **Runs Per Match** plot shows the runs scored by V Kohli in all the matches that he has batted.
- ➢ **Fours and Sixes** plot shows the fours and sixes hit by V Kohli in all the matches that he has batted.

Clearly evident that A Nehra has got the better of V Kohli in 7 matches, i.e. he has dismissed V Kohli the most number of times.
Also, V Kohli has more number of fours than sixes in almost all the matches.

Similar filters can be applied using the **Top Wicket Taker** plot. This change the other plots accordingly.
Also, the user can filter the statistics for any season or match using the Season and Match Id Filter.

Thus, the dashboard is interactive and can be interpreted in many ways as required by the user.

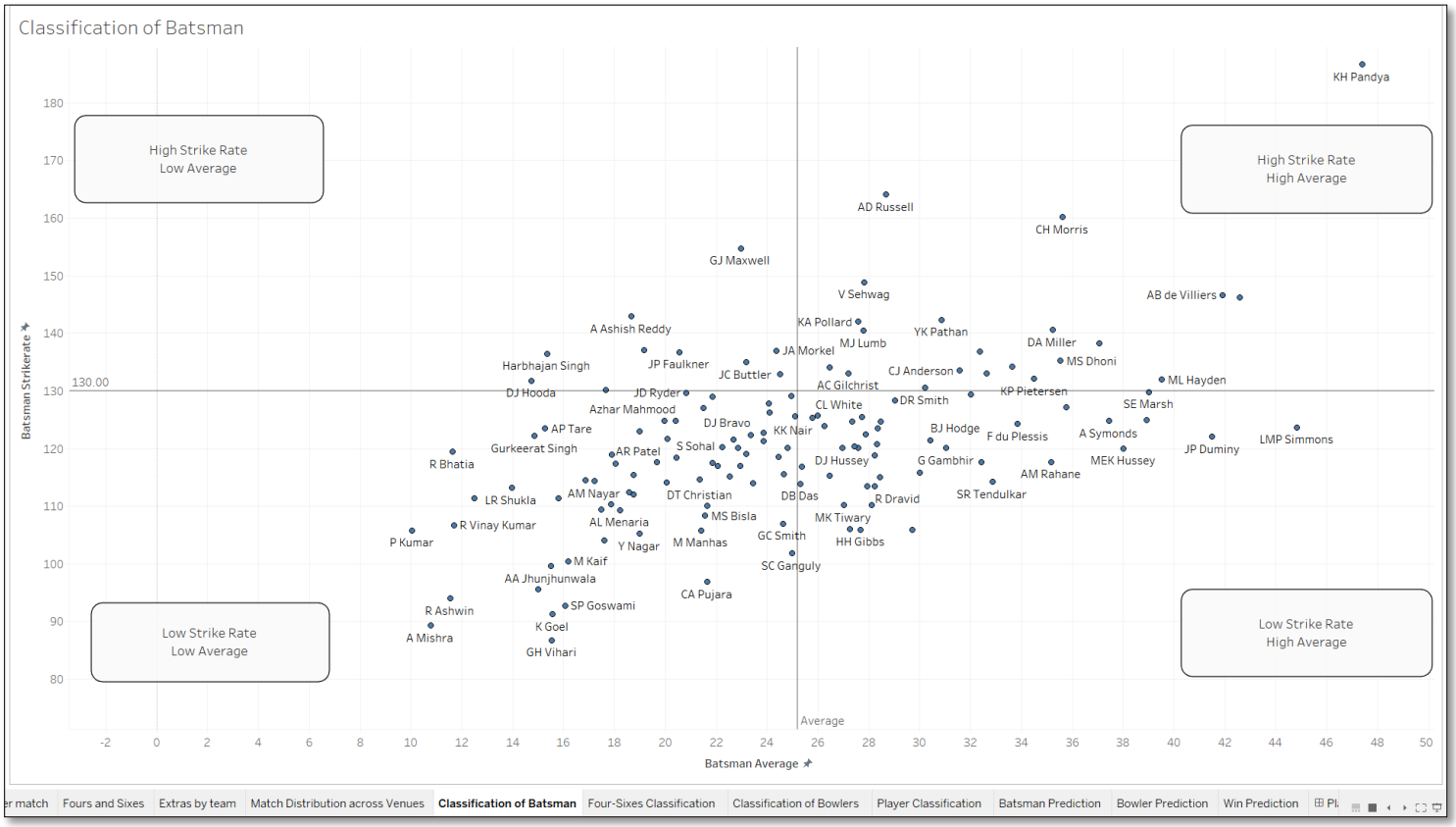
Activity 3: Implement a Data Product

In our Data Product **Cricket_IPL**, we have provided Classification as well as Prediction .

Classification:

Worksheets:

1. **Classification of Batsman:**



Plot (12)

The above **Scatter Plot** (12) classifies the Batsmen into 4 types: High Strike High Average, High Strike Rate Low Average, Low Strike Rate High Average, Low Strike Rate, Low Average.

Fields Used:

1. **Batsman (deliveries.csv):** Provides the list of all the batsmen.

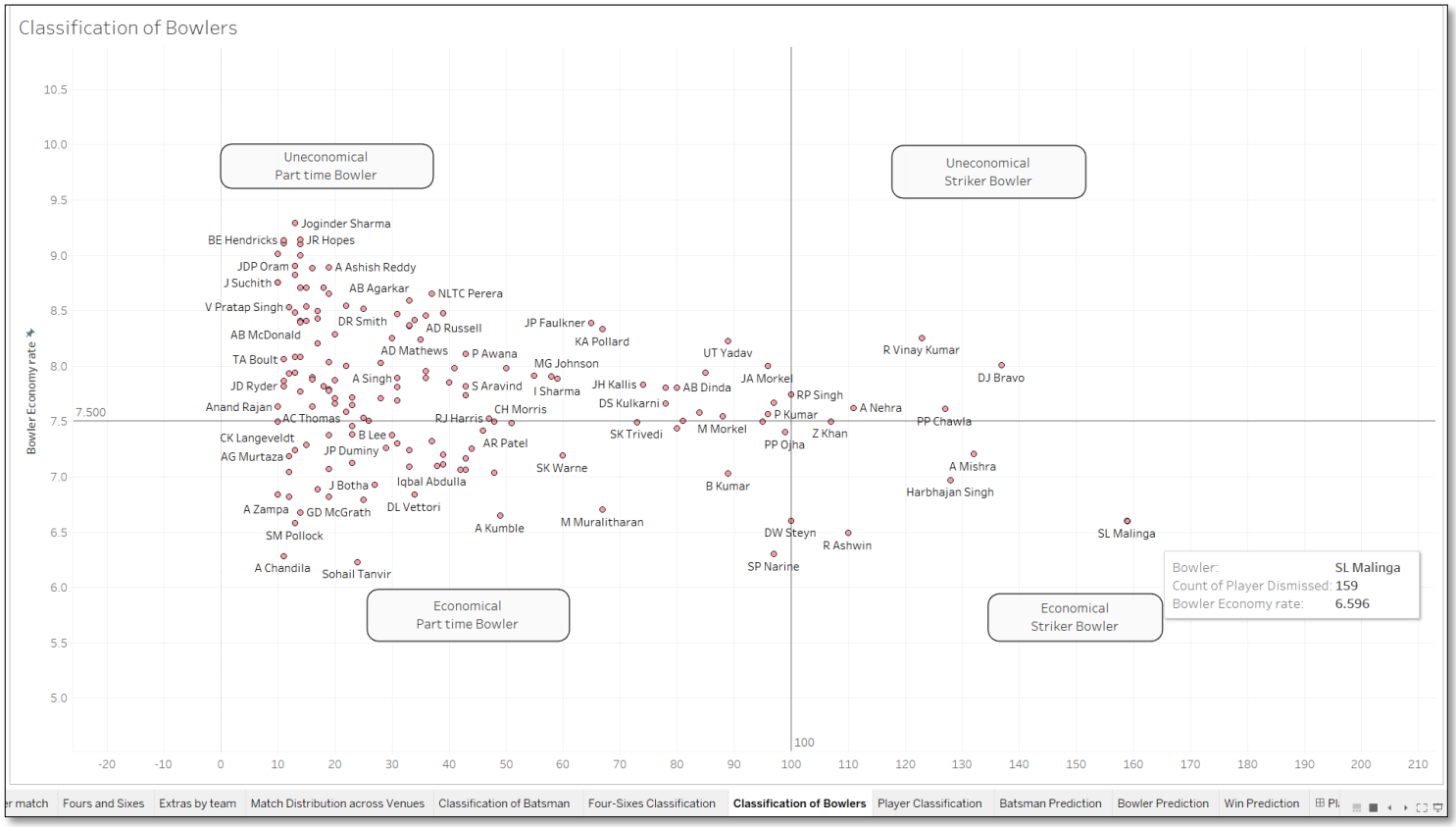
Fields Created:

1. **Batsman Average:** Average for a batsman is defined as the Total number of runs scored divided by the total number of times he has been dismissed in all his match appearances.
   Formula: SUM([Total Runs])/COUNT([Player Dismissed])

2. **Batsman Strike Rate:** Strike Rate for a batsman is defined as the Total number of runs scored every 100 balls.
   Formula: SUM([Batsman Runs])/COUNT([Batsman]) * 100

A batsman in the High Strike Rate High Average area scores huge runs at a healthy pace, while a batsman in the High Strike Rate Low Average area scores at a healthy pace but cannot convert starts into a big innings. Similarly, we can define for the other two areas as well.
Thus, a player in the Right Top area (Average > 26 and Strike rate > 130) is usually considered to be a Hot Property in the IPL due to high strike rate with good average.
Reading the above plot, KH Pandya has the highest average as well as the highest strike rate among all the batsmen.

2. Classification of Bowlers:



Plot (13)

The above **Scatter Plot** (13) classifies the Bowlers into 4 types: Uneconomical Striker Bowler, Economical Striker Bowler, Uneconomical Part time Bowler and Economical Part time Bowler.

Fields Used:

1. **Bowler (deliveries.csv):** Provides the list of all the bowlers.

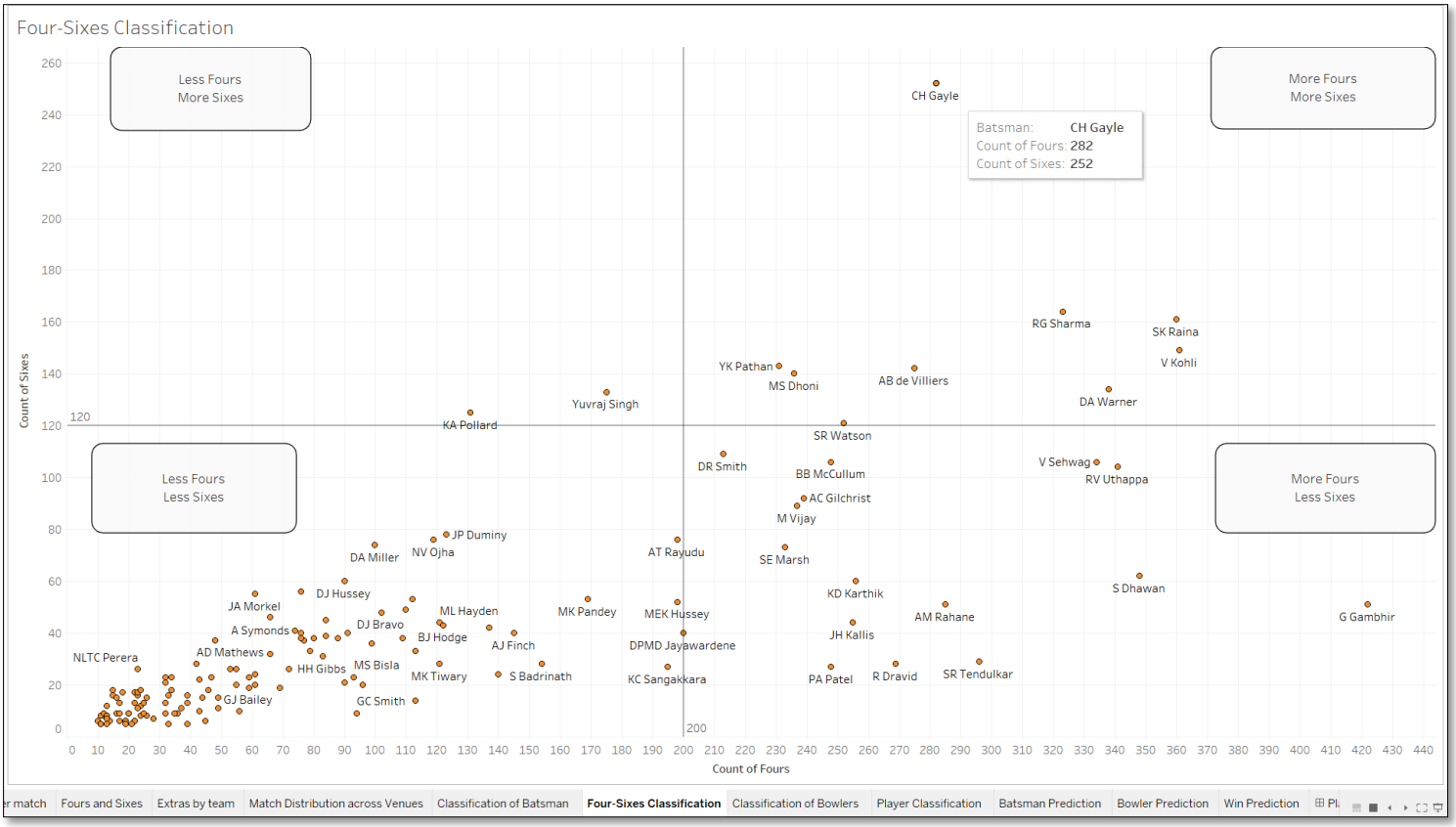
2. **Player Dismissed (deliveries.csv):** Used to count the total number of wickets taken by a bowler.

Fields Created:

1. **Bowler Economy Rate:** Economy rate for a bowler is defined as the average number of runs conceded by the bowler in an over.
   Formula: SUM([Total Runs])/COUNT([Bowler]) *6

A bowler in the Uneconomical Striker Bowler area provides strike throughs but also leaks away runs at a good pace, while a bowler in the Economical part time Bowler area may not take as many wickets but is good at ceasing the flow of runs.
Thus, a player in the Right Bottom area (Economy rate < 7.5 and Number of wickets >100) is usually considered to be a good bowler in the shorter format of the game due to low economy rate and high capability of providing wickets.
Reading the above plot, SL Malinga has a decent economy rate with the most number of wickets.

3. Fours-Sixes Classification:



Plot (14)

The above **Scatter Plot** (14) classifies the batsmen based on the number of boundaries and sixes hit: Less Fours Less Sixes, More Fours More Sixes, Less Fours More Sixes and More Fours and Less Sixes.

Fields Used:

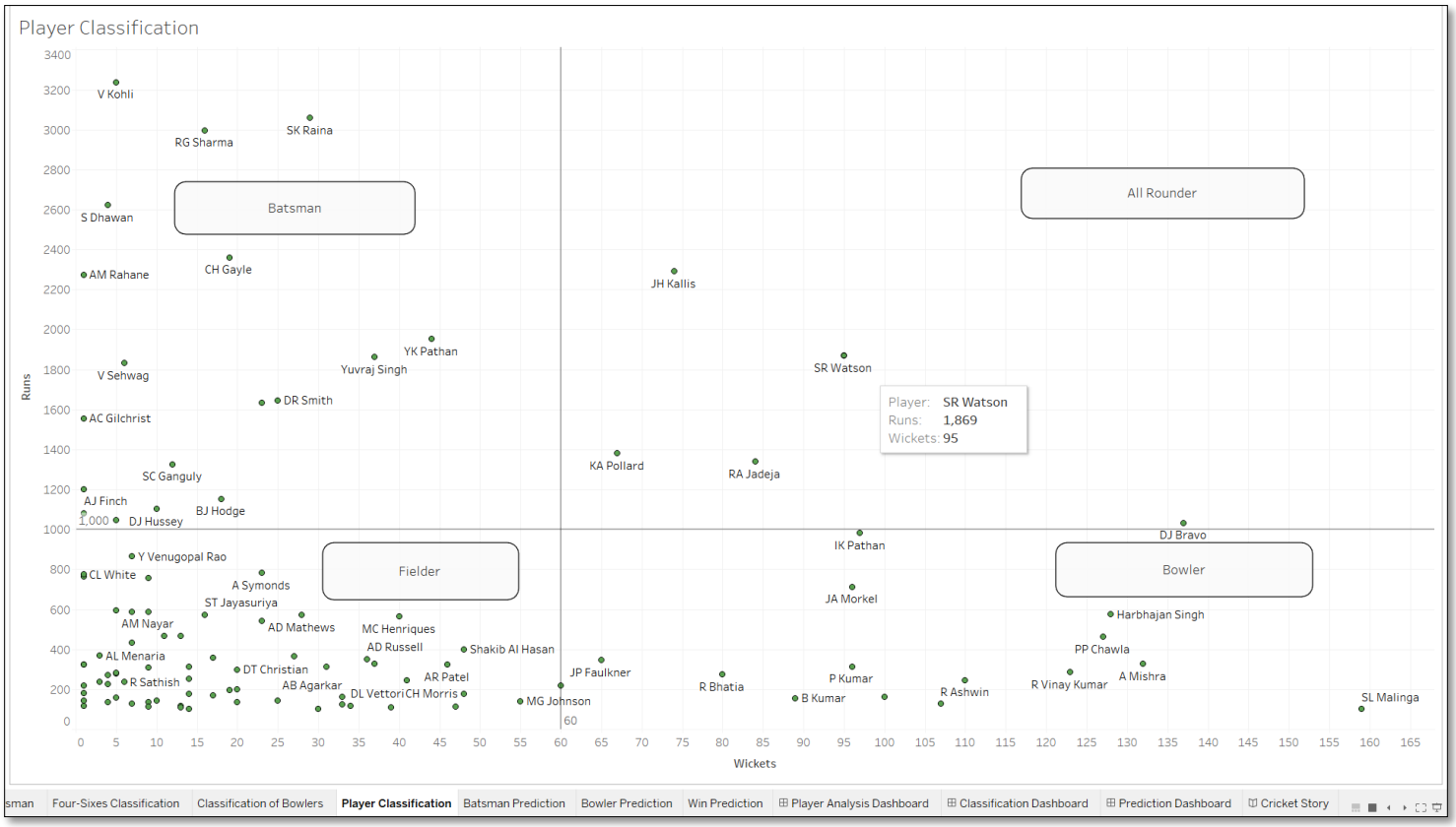
1. **Batsman (deliveries.csv):** Provides the list of all the batsmen.

2. **Fours (created):** Counts the number of fours scored by a batsman.

3. **Sixes (created):** Counts the number of sixes scored by a batsman.

A batsman in the More Sixes More Fours area hits as many number of sixes as number of fours while a batsman in the Less Sixes More Fours hits more boundaries than sixes.
Thus, a player in the Right top area (Count of Sixes > 120 and Count of Fours > 200) is usually liked by the audiences as he provides more entertainment by hitting a lot of fours and equal number of sixes.
Reading the above plot, Chris Gayle has the most number of sixes with almost equal number of fours.

4. Player Classification:



Plot (15)

The above **Scatter Plot** (15) classifies the players into 4 types: All rounder, Batsman, Bowler and Fielder. This is based on the number of runs scored and the number of wickets taken by a player.
We have used R-code for calculating the total number of runs scored and total number of wickets by a Player and written the same in a csv file Statistics.csv.
We have created a new Data Source **Statistics** using this csv file

Fields Used:

1. **Player (Statistics.csv):** Provides the list of all Players to have played the IPL.

2. **Runs (Statistics.csv):** Aggregates to give the total number of runs scored by a player.

3. **Wickets (Statistics.csv):** Aggregates to give the total number of wickets taken by a player.

A player in the All Rounder area generally the one with high runs and high wickets, i.e. he performs well in both the departments of the game while a player in the Batsman area is the one with more runs and less/no wickets. Such a type of player is generally the top order batsman of a team. Similarly, a player in the Bowler area is the one with more wickets and very less runs. Such a type of player is generally the front line bowler of a team.
Reading the plot above, players like SR Watson and JH Kallis are good all rounders of the game while players like V Kohli and RG Sharma are specialist batsmen.

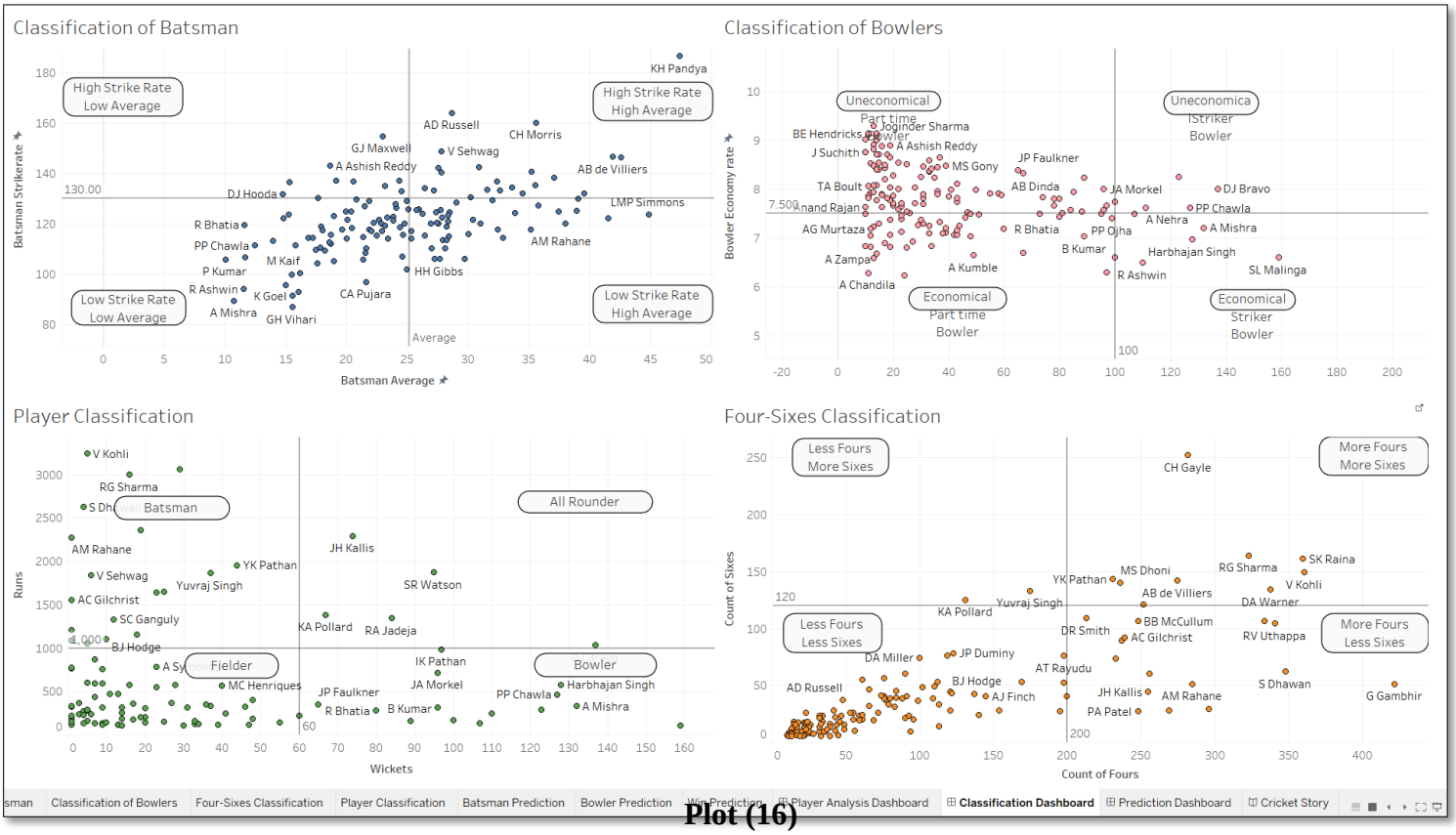
Dashboard:

Classification Dashboard:

The below plot (16) provides a dashboard which classifies the Players, Batsmen, Bowlers and runs scored by fours and Sixes.
It is the collection of all the classification worksheets described above. A user can easily check the classifications of all the Players at one place using this dashboard.

Worksheets used:

1. **Classification of Batsman**
2. **Classification of Bowlers**
3. **Four-Sixes Classification**
4. **Classification of Players**



Plot (16)

Using this dashboard, we can classify the Role of a Player in a team. Roles can vary from a Main Bowler to an Opening batsman or a Pinch Hitter to a finisher.

If a player falls in the Batsman part of the **Player Classification** plot and has high strike rate and high average as per the **Classification of Batsman** plot, then he maybe the best choice for opening the innings for a team. Similarly, for a player in the Bowler part of the **Player Classification** plot who is an **economical striker bowler** as per the Classification of Bowler plot can be the front line bowler for any team providing early breakthroughs and not leaking too many runs.
Likewise, we can predict the role of a player for the other combinations as well.

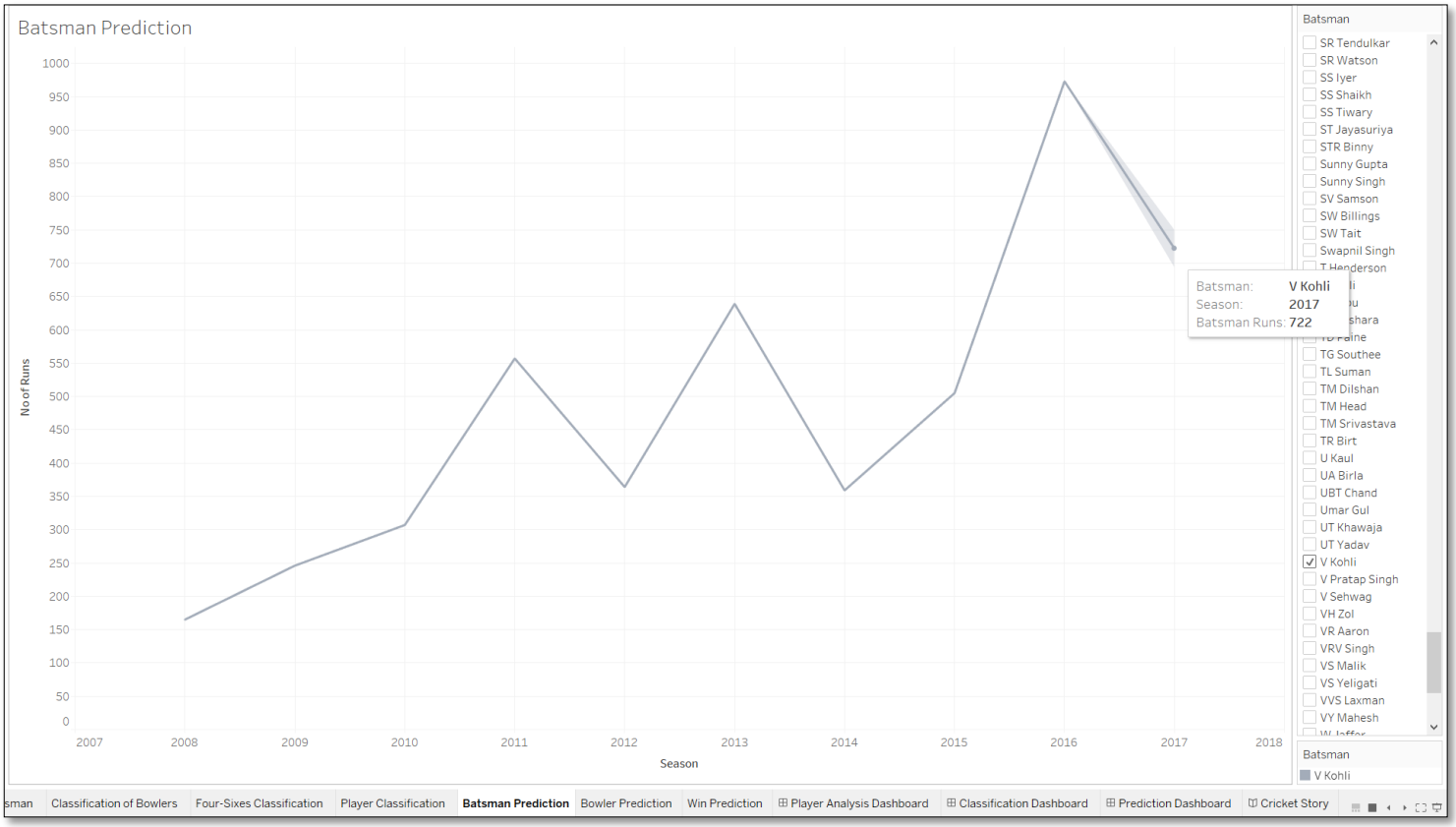
Prediction:

Forecast:

- We have used the Forecast feature of tableau for predictions using old data.
- Forecasting in Tableau uses a technique known as exponential smoothing. Forecast algorithms try to find a regular pattern in measures that can be continued into the future.
- We can typically add a forecast to a view that contains a date field and at least one measure.

Worksheets:

1. Batsman Prediction:



Plot (17)

The above **Line Plot** (17) shows the runs scored by a player across different seasons of the IPL. We have the data for IPL from 2008-2016. Using the **Forecast** feature of the Tableau, we are calculating the number of runs the batsman is likely to score in 2017.

Fields Used:

1. **Batsman Runs (deliveries.csv):** Aggregated to get the total number of runs scored by a batsman.

2. **Season (matches.csv):** Provides the list of all the seasons the IPL has been played in.

3. **Batsman (deliveries.csv):** Provides a list of all the batsman. Also, the plot is colored based on the batsman field.

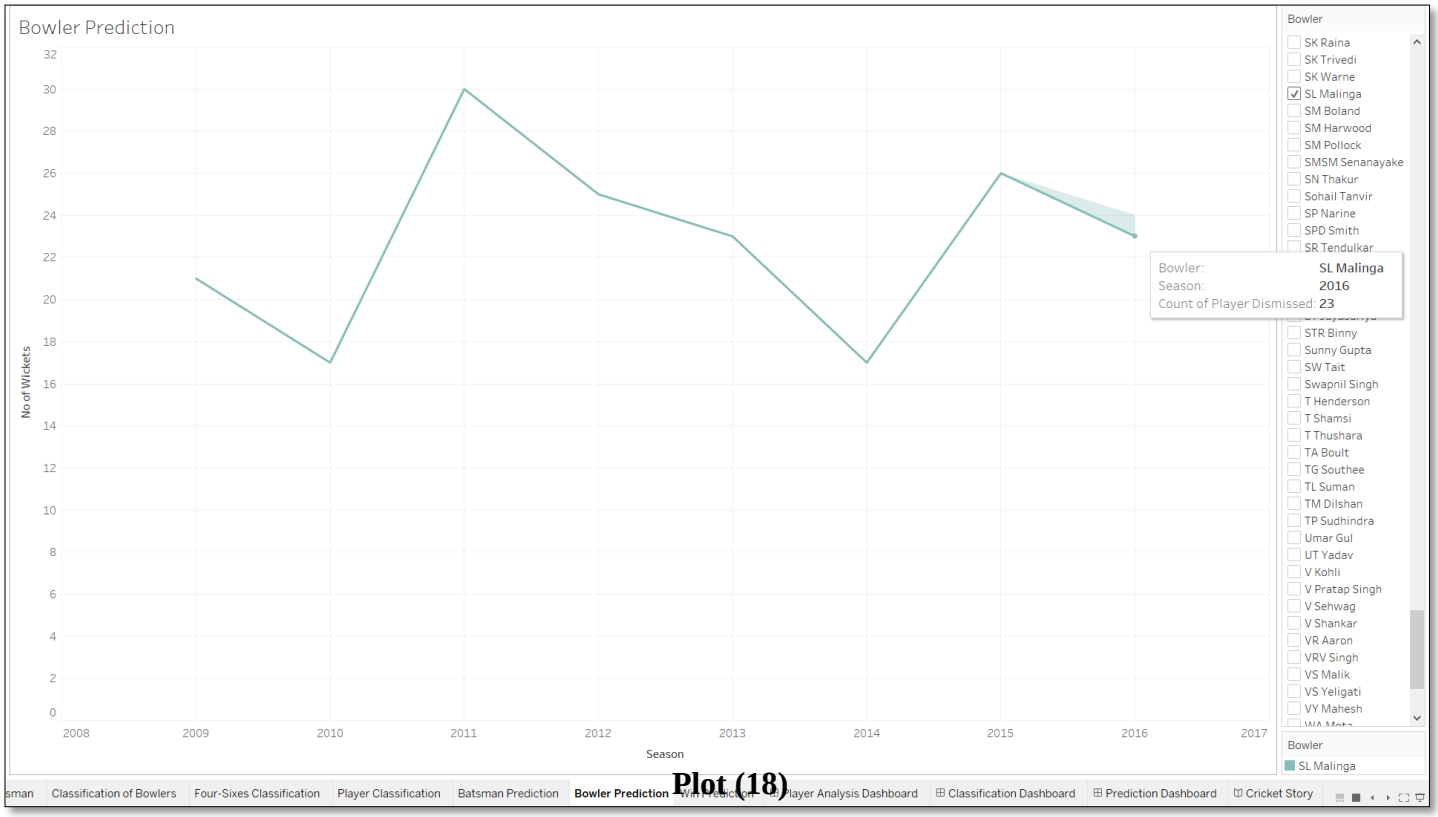
Below is the forecast model used by the Tableau in this case (e.g. for Batsman V Kohli):

All forecasts were computed using exponential smoothing.

Sum of Batsman Runs

Color	Model			Quality Metrics					Smoothing Coefficients			
Batsman	Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma	
V Kohli	Additive	None	None	224	170	0.78	36.40%	103	0.5	0		0

2. Bowler Prediction:



Plot (18)

The above **Line Plot** (18) shows the wickets taken by a player across different seasons of the IPL. We have the data for IPL from 2008-2016. Using the **Forecast** feature of the Tableau, we are calculating the number of wickets the bowler is likely to take in 2017.

Fields Used:

1. **Player Dismissed (deliveries.csv):** It gives the count of the total number of wickets taken by a bowler.

2. **Season (matches.csv):** Provides the list of all the seasons the IPL has been played in.

3. **Bowler (deliveries.csv):** Provides a list of all the bowlers. Also, the plot is colored based on the bowler field.

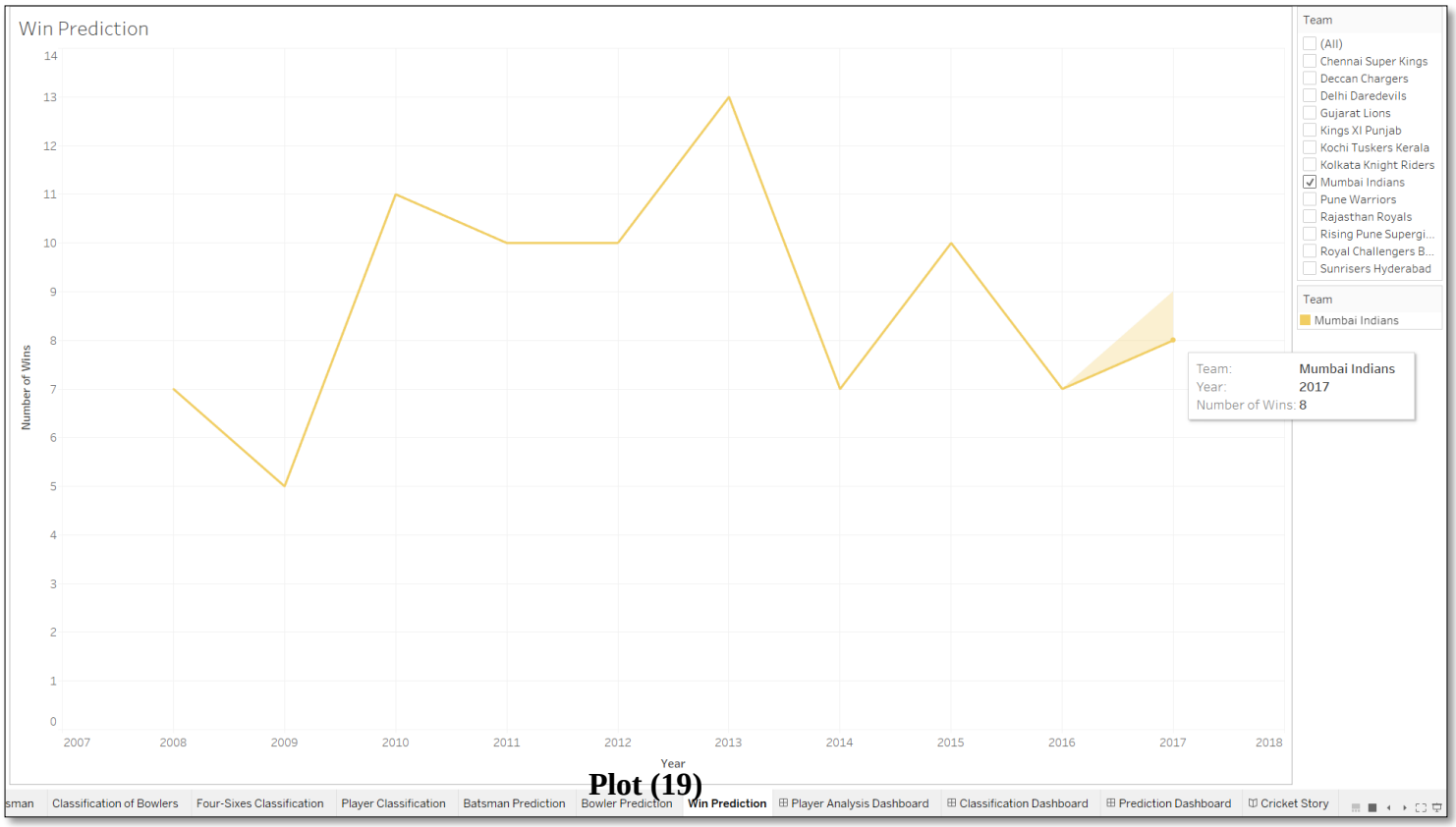
Below is the forecast model used by the Tableau in this case (e.g. for Bowler SL Malinga):

All forecasts were computed using exponential smoothing.

Count of Player Dismissed

Color	Model			Quality Metrics					Smoothing Coefficients		
Bowler	Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
SL Malinga	Additive	None	None	4	4	0.58	17.90%	27	0	0	0

3. Win Prediction:



Plot (19)

The above Line Plot (19) shows the number of wins for each team across different seasons of the IPL. Using the **Forecast** feature of tableau, we are predicting the number of games a team is likely to win in the coming season of the tournament.

We have used R-code for calculating the total number of wins per team per year and written the same in a csv file Winner.csv.
We have created a new Data Source **Wins** using this csv file

Fields Used:

1. **Year (Winner.csv):** Provides the list of all the seasons the tournament was played.
2. **Number of wins (Winner.csv):** Gives the total number of wins for a team for each year.
3. **Team (Winner.csv):** Provides the list of all the teams. Also, used to color the plot.

Below is the forecast model used by the Tableau in this case (e.g. for Team Mumbai Indians):

All forecasts were computed using exponential smoothing.

Sum of Number of Wins

Color	Model			Quality Metrics					Smoothing Coefficients		
Team	Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Mumbai Indians	Additive	None	None	2	2	0.73	25.80%	22	0	0	0

Dashboard:

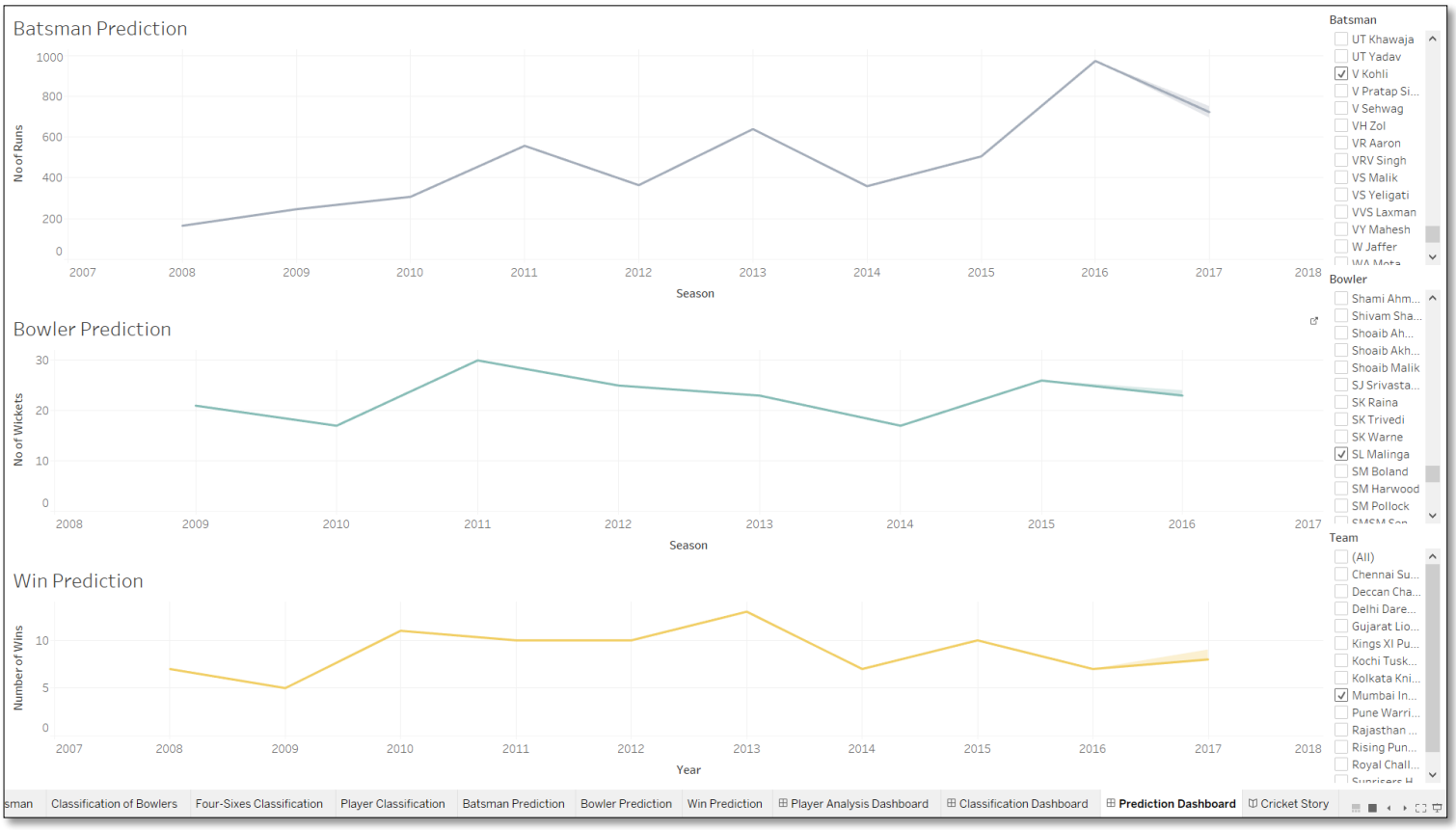
Prediction Dashboard:

The below plot (20) shows a dashboard which can be used to Predict for a Batsman, Bowler and a Team, all at the same page.
It is the collection of all the prediction worksheets described above.

Worksheets Used:

1. **Batsman Prediction**
2. **Bowler Prediction**
3. **Win Prediction**

Using this dashboard, one can predict the performance of a player in the coming season of the tournament. This prediction can be used by the Team Managers to buy/sell the players at an increased or decreased price in the auctions.
Also, the dashboard can be used to predict the performance of a team in the coming season. This prediction can be used to find the odds of a team winning the tournament.


Batsman Prediction
No of Runs
Season
Bowler Prediction
No of Wickets
Season
Win Prediction
Number of Wins
Year
Batsman
UT Khawaja
UT Yadav
V Kohli
V Pratap Si...
V Sehwag
VH Zol
VR Aaron
VRV Singh
VS Malik
VS Yeligati
VVS Laxman
VY Mahesh
W Jaffer
Bowler
Shami Ahm...
Shivam Sha...
Shoaib Ah...
Shoaib Akh...
Shoaib Malik
SJ Srivasta...
SK Raina
SK Trivedi
SK Warne
SL Malinga
SM Boland
SM Harwood
SM Pollock
Team
(All)
Chennai Su...
Deccan Cha...
Delhi Dare...
Gujarat Lio...
Kings XI Pu...
Kochi Tusk...
Kolkata Kni...
Mumbai In...
Pune Warri...
Rajasthan ...
Rising Pun...
Royal Chall...
sman
Classification of Bowlers
Four-Sixes Classification
Player Classification
Batsman Prediction
Bowler Prediction
Win Prediction
Player Analysis Dashboard
Classification Dashboard
Prediction Dashboard
Cricket Story

Plot (20)

Story:

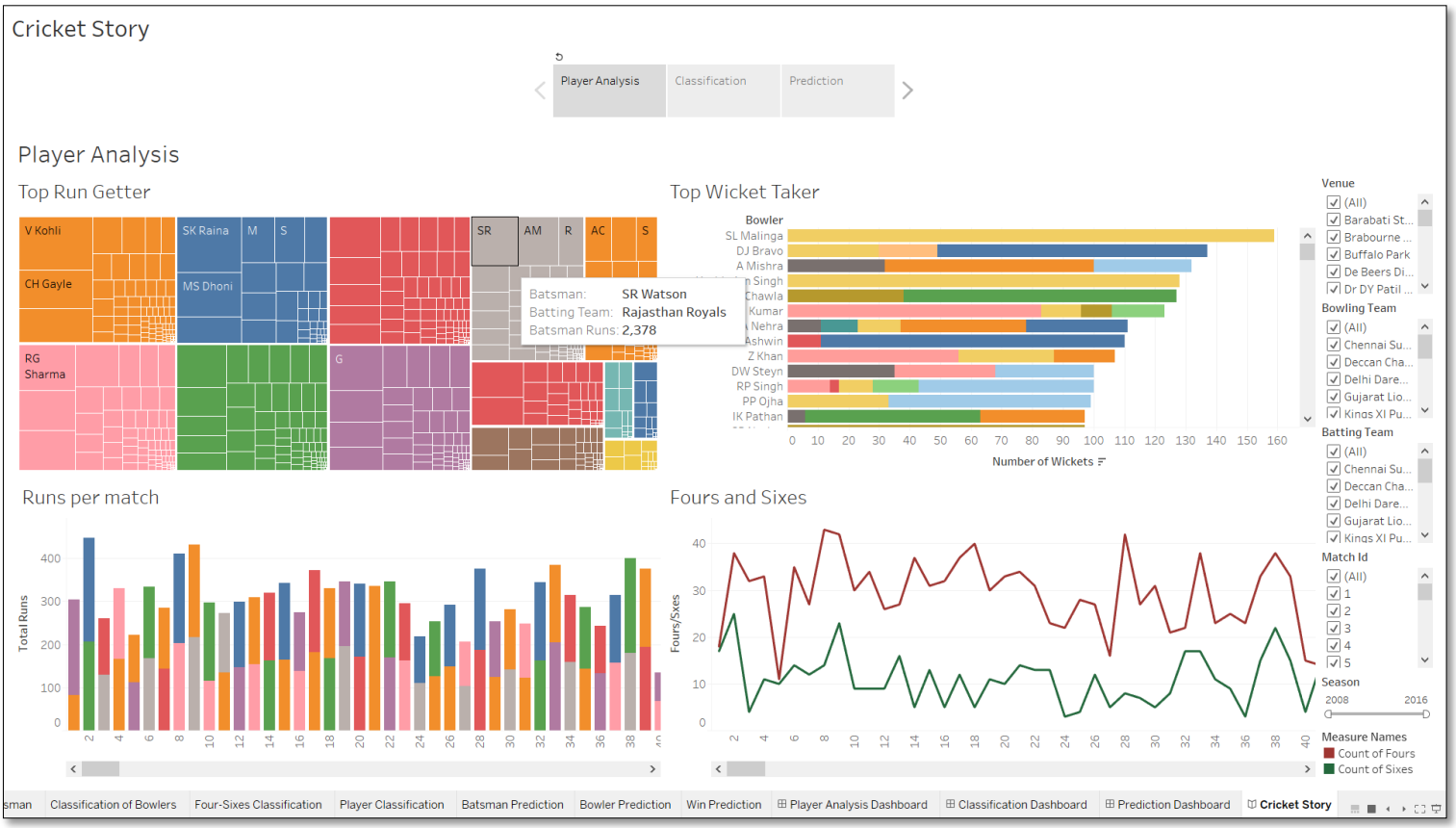
Cricket Story:

This is a sequence of all the Dashboards created above. The user can check the Story instead of checking all the datasheets or dashboards individually.

Dashboards Used:

1. **Player Analysis Dashboard**
2. **Classification Dashboard**
3. **Prediction Dashboard**

Users can navigate through different dashboards and check the functionalities each dashboard has to offer.



Plot (21)

Tableau Link:

https://public.tableau.com/profile/saurabh.bajoria#!/vizhome/Cricket_IPL_0/TopRungetter

References:

- http://onlinehelp.tableau.com/current/pro/desktop/en-us/stories.html
- http://onlinehelp.tableau.com/current/pro/desktop/en-us/forecast_how_it_works.html
- https://www.tutorialspoint.com/tableau/tableau_dashboard.htm
- https://www.kaggle.com/manasgarg/ipl