

Indian Premier League 2008-2019

EDA and regression

By: Sanjay jaras

Date: 05/30/2020



Introduction

- **Indian Premier League 2008-2019**

The Indian Premier League (IPL) is a professional Twenty20 cricket league in India. It is typically contested during March through May of every year by eight teams representing eight different cities in India. The Board of Control for Cricket in India (BCCI) started the league in 2008. The IPL has a timeslot in ICC (International Cricket Council) Future Tours Programme. The IPL is the most-attended cricket league in the world. In 2014 IPL is ranked sixth by average attendance among all sports leagues. The brand value of the IPL in 2019 was ₹475 billion (US\$6.7 billion), according to Duff & Phelps. According to BCCI, the 2015 IPL season contributed ₹11.5 billion (US\$160 million) to the GDP of the Indian economy.

The dataset I got from Kaggle.com has ball by ball information of all seasons of IPL that happened from 2008 through 2019. This dataset contains a total of 751 matches played. This dataset contains one yaml file for each game. This data needs to be used combinedly for performing analysis; for this, we need to combine data from all yaml files.

I have tried to answer different questions by using this dataset.

Data source:

<https://www.kaggle.com/sagara9595/indian-premier-league-20082019>

<https://cricsheet.org/>

Statistical/Hypothetical Questions

Does Virat Kohli score more run in second innings?

Analytical distribution comparison

Which batsman has best strike rate?

Is average score per innings increasing by season?

Is Mumbai-Indians better team?

Does runs scored in first 6 overs or last 5 overs decided match result?

Does runs given in first 6 overs or last 5 overs decided match result?

Does opener batsmen's contribution decide match results?

Predict match result by using different features.

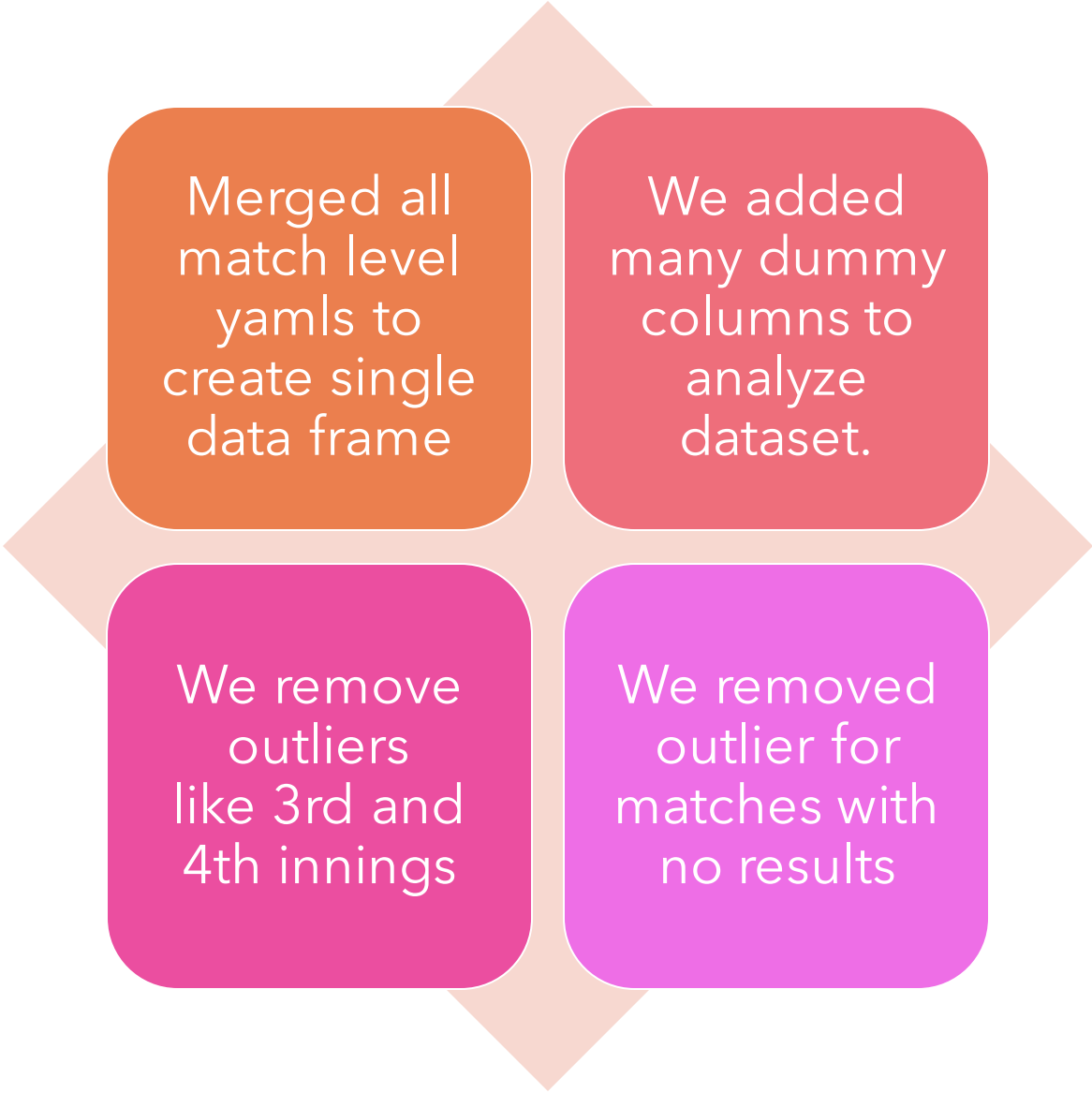
Which player got more centuries and half-centuries?

Which team won more tosses ?

Which team won more matches ?

Which player got more man-of-the-match awards?

Data Preparation



Merged all
match level
ymls to
create single
data frame

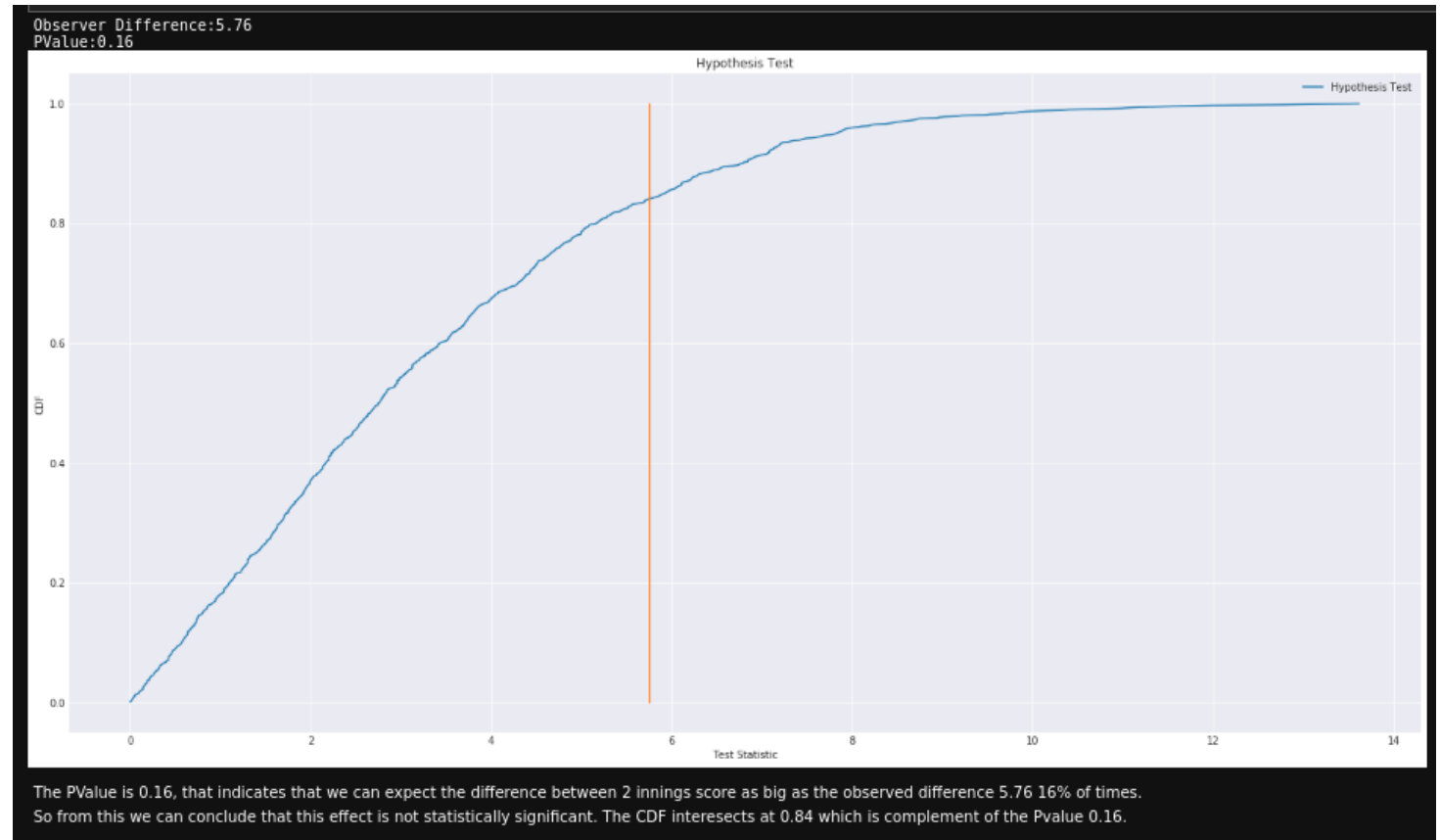
We added
many dummy
columns to
analyze
dataset.

We remove
outliers
like 3rd and
4th innings

We removed
outlier for
matches with
no results

Does Virat Kohli score more run in second innings?

- After comparing scores by using hypothesis testing for Virat Kohli.
- The PValue is 0.16, that indicates that we can expect the difference between 2nd innings score as big as the observed difference 5.76 16% of times.
- So from this we can conclude that this effect is not statistically significant. The CDF intersects at 0.84 which is complement of the PValue 0.16.

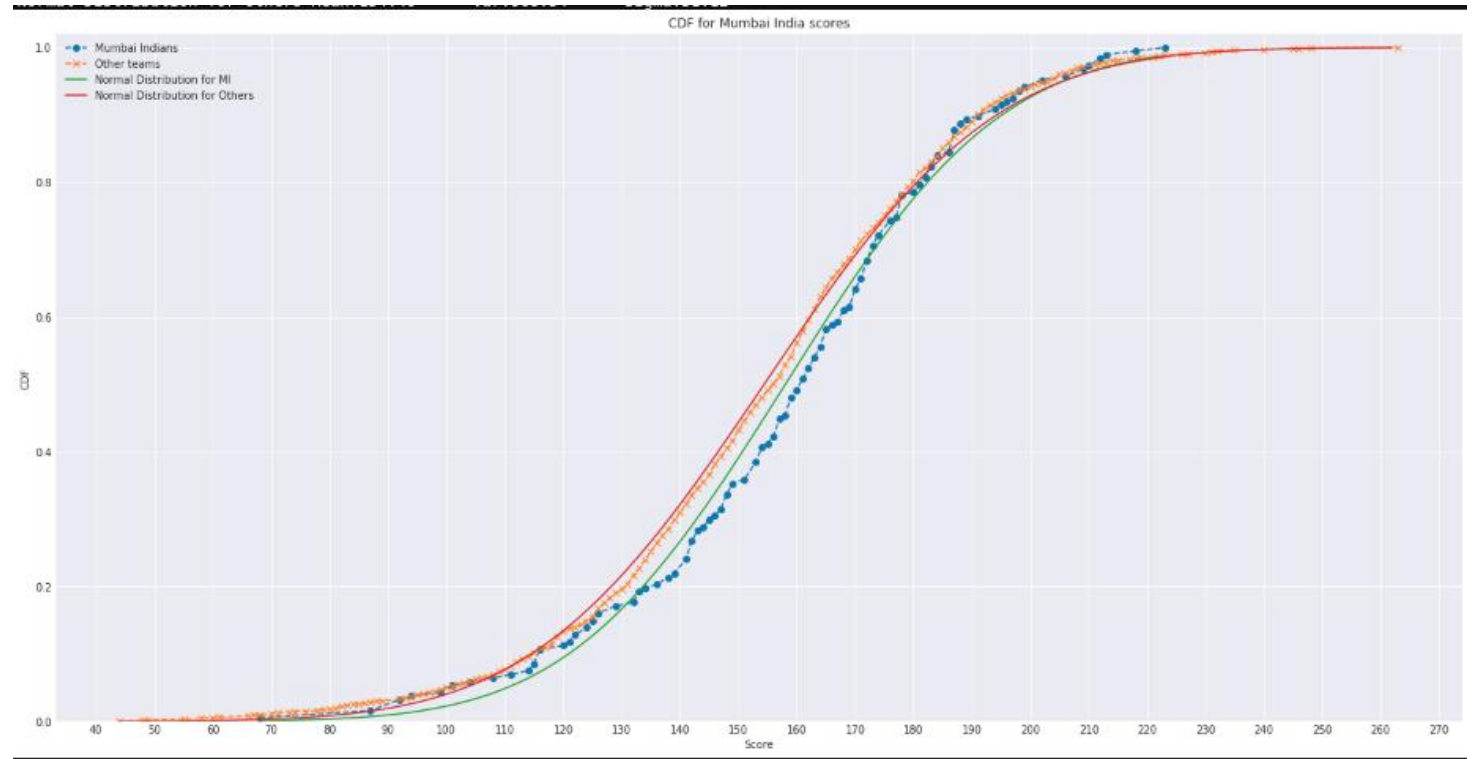


Analytical Distribution Comparison

The team scores CDF are matching with normal distribution CDF.

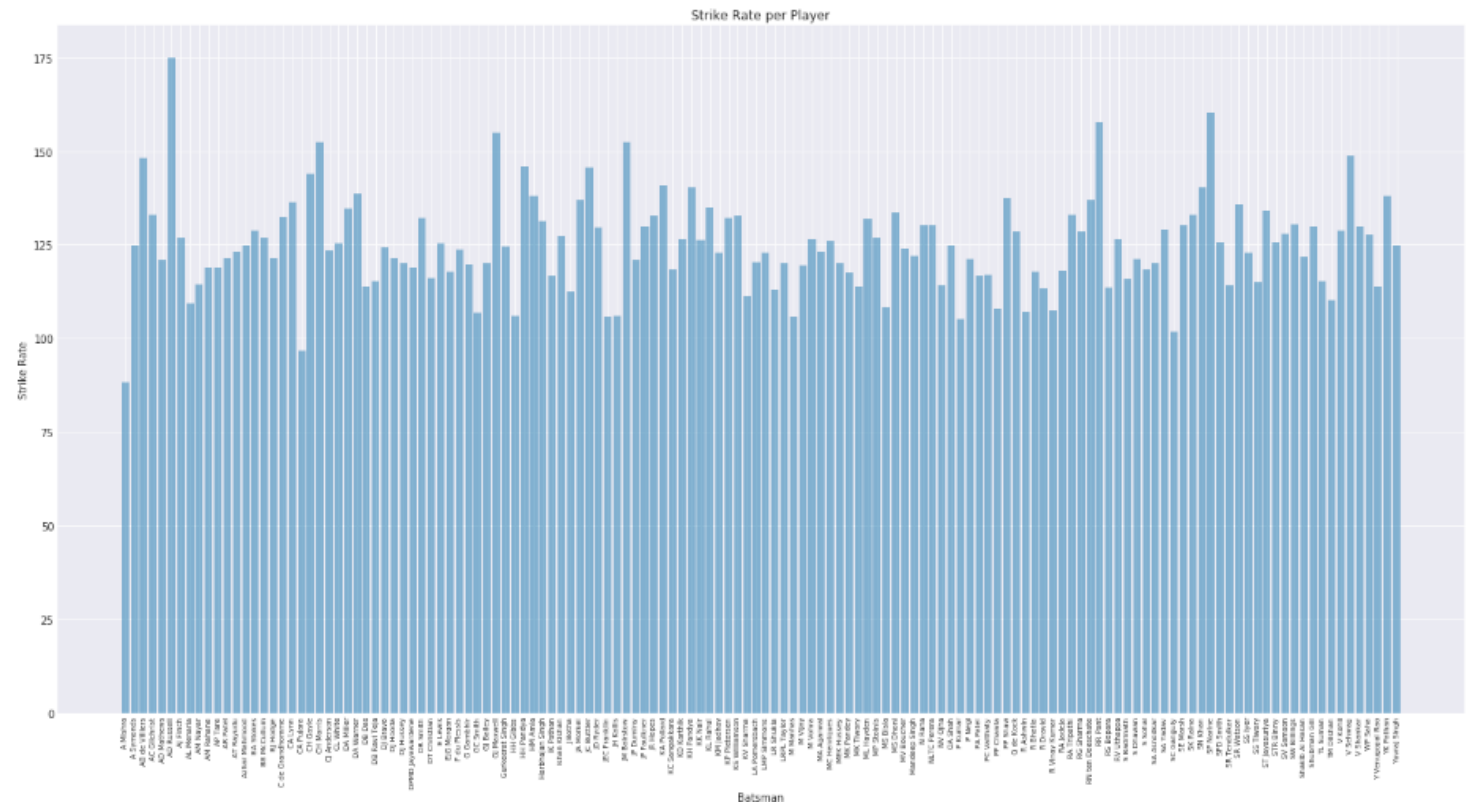
Normal Distribution for MI
Mean:158.03 Var:838.68
Sigma:28.96

Normal Distribution for Others
Mean:154.40 Var:968.54
Sigma:31.12



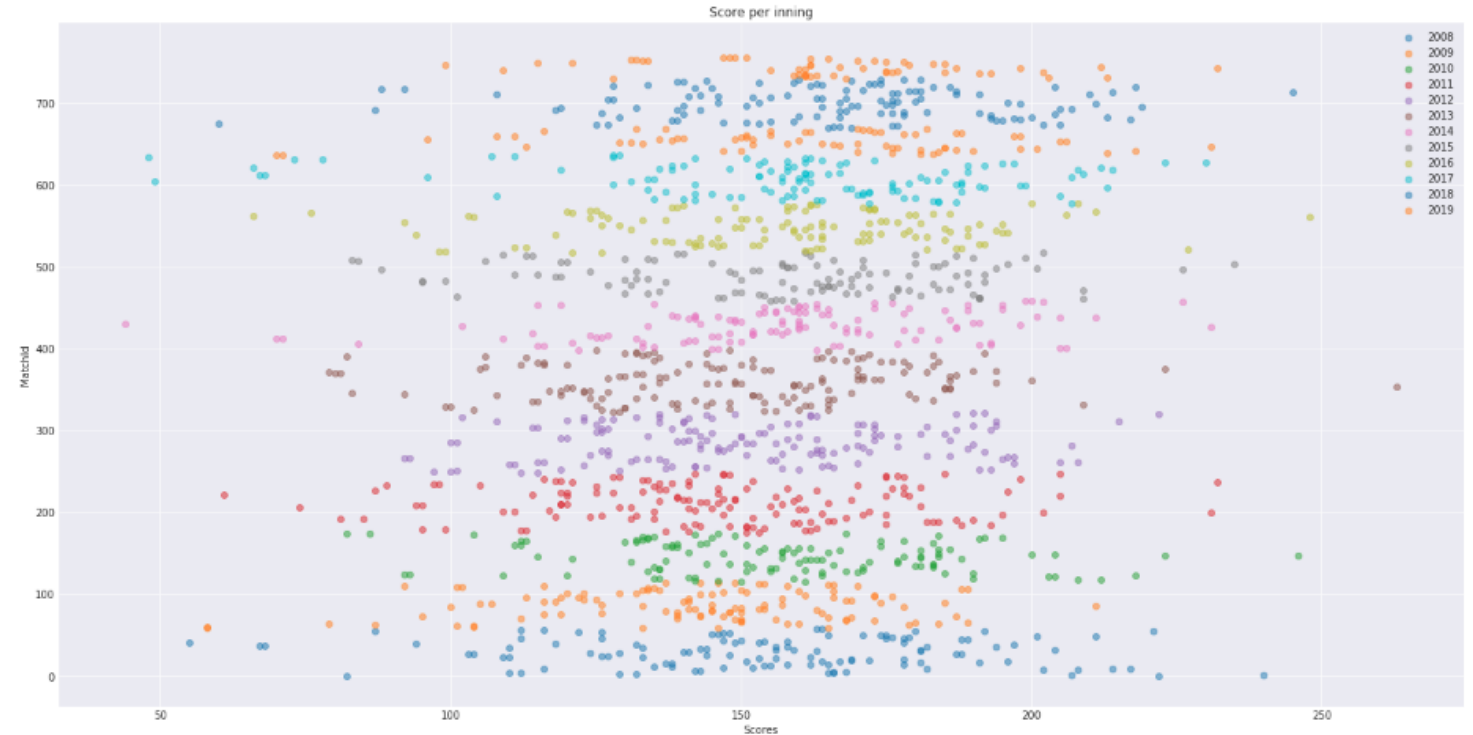
Which batsman has best strike rate?

- From the histogram we can conclude that Aundre Russel having better strike rate 175.



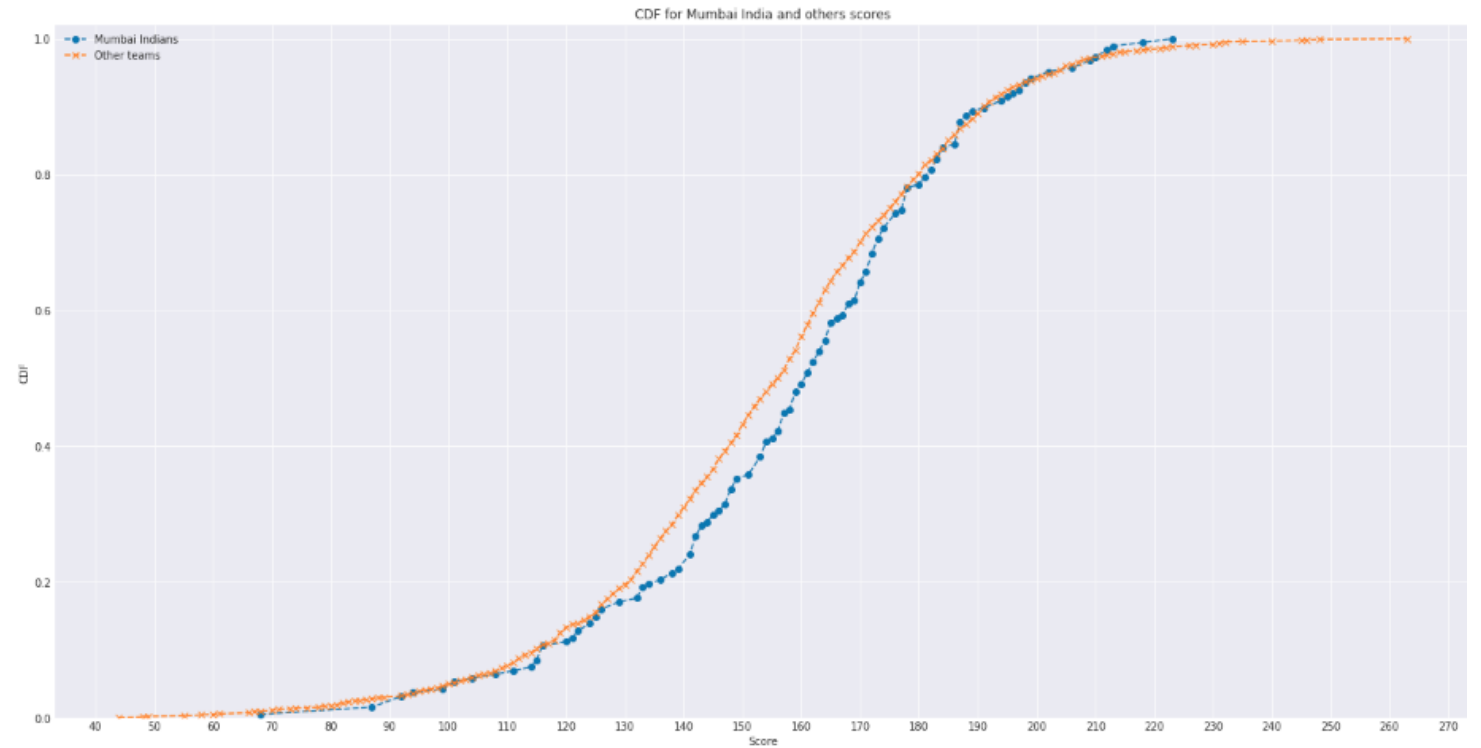
Is average score per innings increasing by season?

- The Scatter plot and the mean by season shows average score by increasing with time.
- Season
- 2008 154.629310
- 2009 143.157895
- 2010 157.200000
- 2011 146.513889
- 2012 151.709459
- 2013 148.296053
- 2014 157.575000
- 2015 157.394737
- 2016 157.183333
- 2017 159.059322
- 2018 165.841667
- 2019 163.533898



Is Mumbai-Indians better team?

- The comparison CDF shows Mumbai Indians has better scores around 130 to 185.



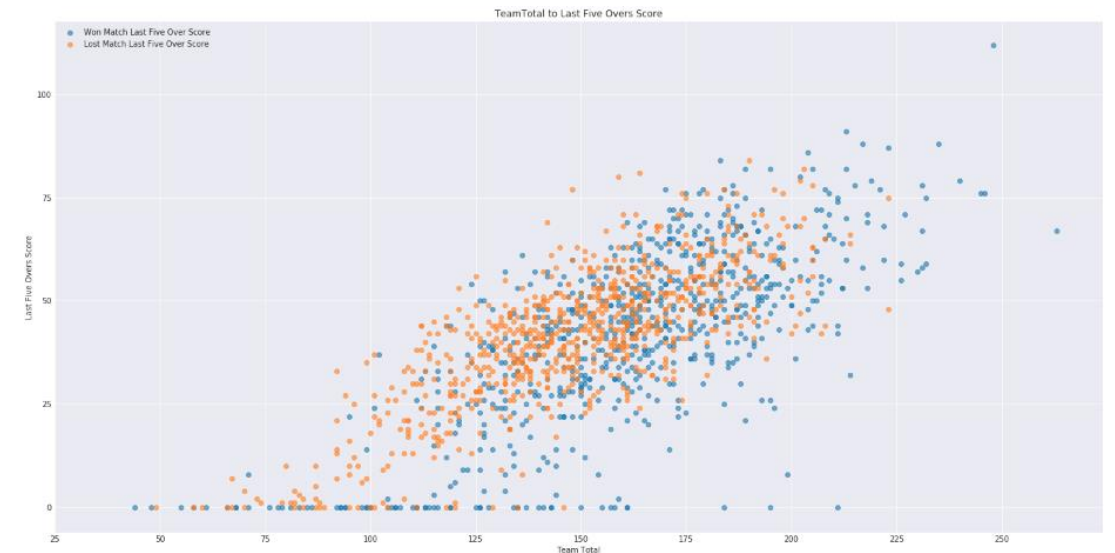
Does runs scored in first 6 overs or last 5 overs decided match result?

• First 6 Overs

- Covariance teamTotalRuns firstSixTotal
- teamTotalRuns 954.46177 147.381050
- firstSixTotal 147.38105 145.055839
- Pearson Correlation teamTotalRuns firstSixTotal
- teamTotalRuns 1.000000 0.396091
- firstSixTotal 0.396091 1.000000
- PointbiseriaIrResult(correlation=0.2283302908772324, pvalue=3.066031954648354e-19)
- Above corretion numbers show we have positive relationship between first six overs score and total team score. Point biserial Correlation values shows positive relationship as well p-value rejects null hypothesis.

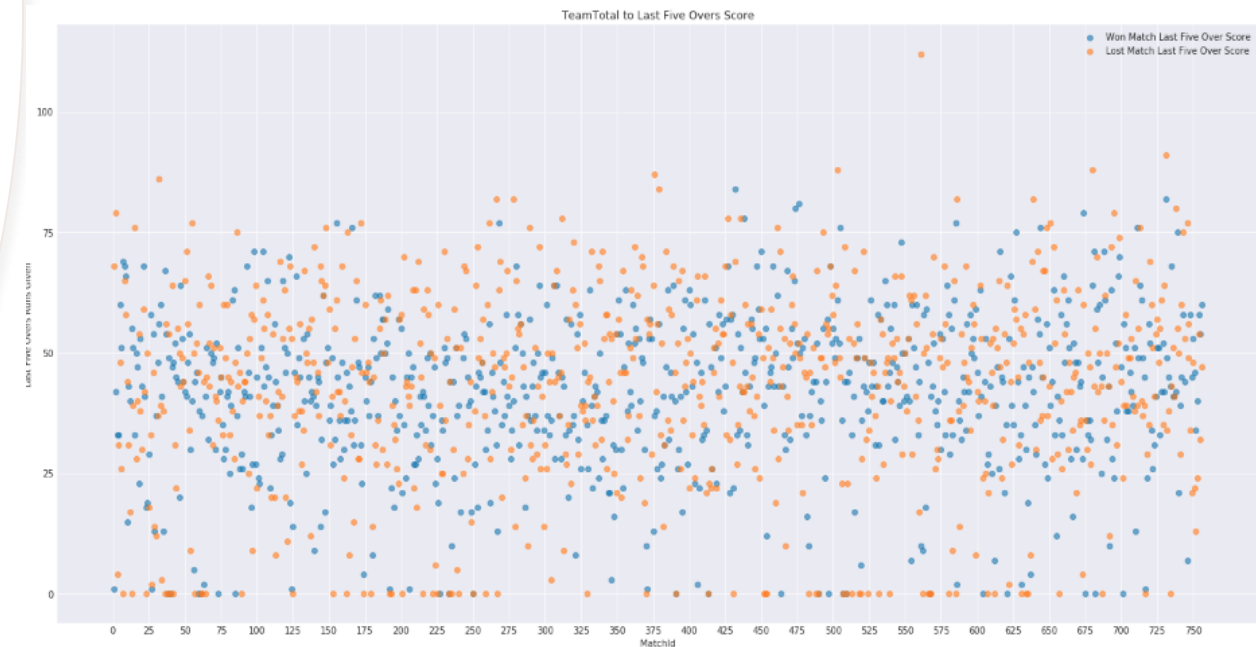
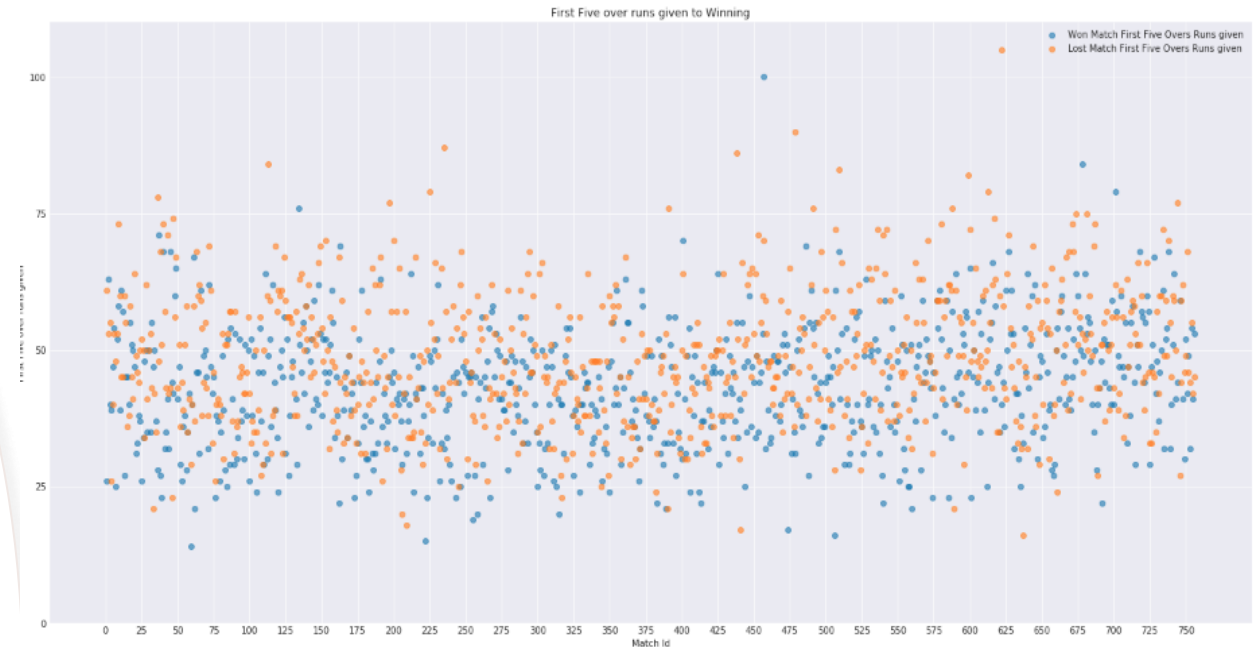
• Last 5 Overs

- Covariance teamTotalRuns lastFiveTotal
- teamTotalRuns 954.461770 418.696843
- lastFiveTotal 418.696843 348.678816
- Pearson Correlation teamTotalRuns lastFiveTotal
- teamTotalRuns 1.000000 0.725784
- lastFiveTotal 0.725784 1.000000



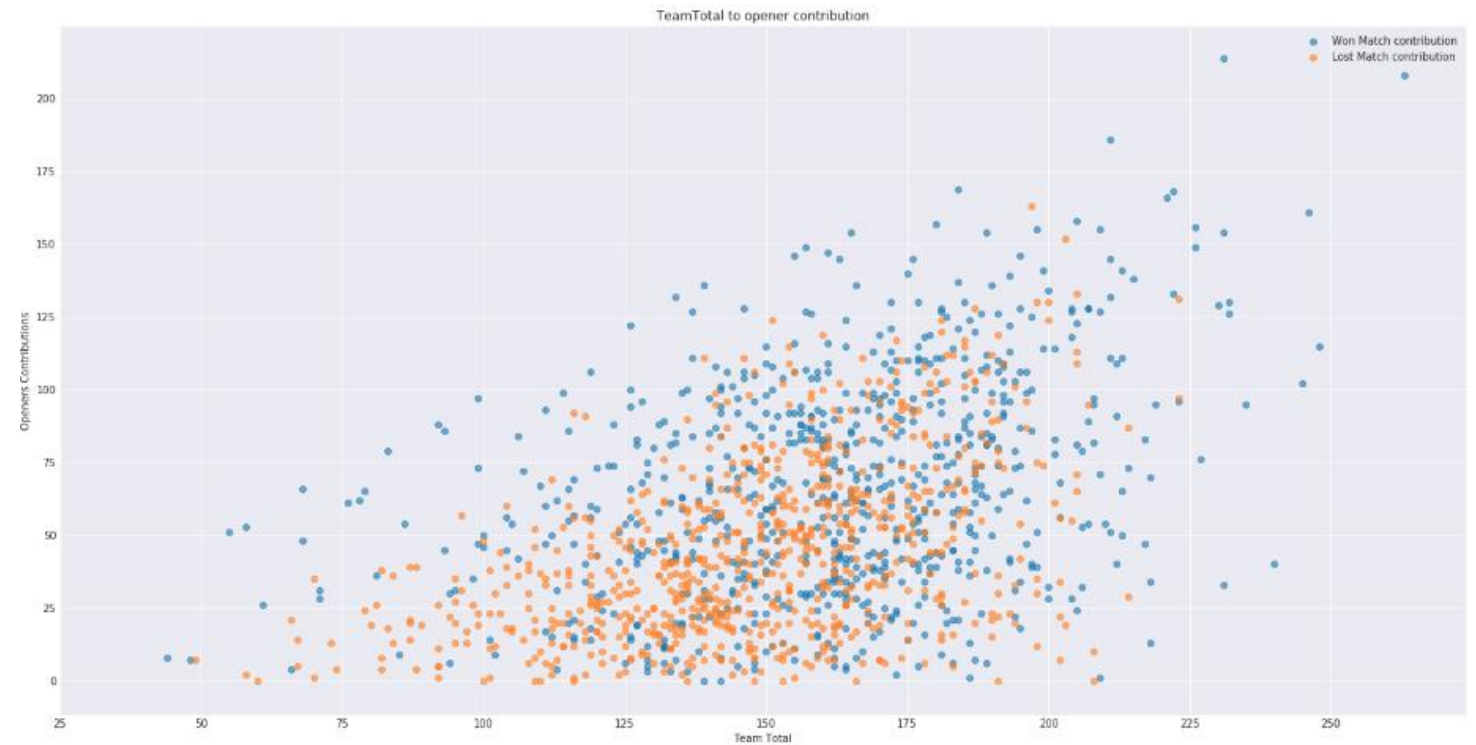
Does runs given in first 6 overs or last 5 overs decided match result?

- `PointbiserialResult(correlation=-0.22498194053040915, pvalue=1.0344301557061923e-18)`
- Above correlation numbers show we have negative relationship between first six overs runs given to opponent and winning and losing. P-value is rejecting null hypothesis as p-value is very small.
- `PointbiserialResult(correlation=-0.03338726466980052, pvalue=0.19563420150438063)`
- Above numbers shows that the relationship is very weak also p-value indicates the null hypothesis is not false



Does opener batsmen's contribution decide match results?

- Covariance teamTotalRuns OpenersTotalRuns
- teamTotalRuns 954.461770 501.838433
- OpenersTotalRuns 501.838433 1266.492625
- Pearson
Correlation teamTotalRuns OpenersTotalRuns
- teamTotalRuns 1.00000 0.45644
- OpenersTotalRuns 0.45644 1.00000
- PointbiserialrResult(correlation=0.3064424433593573,
pvalue=4.583588404215659e-34)
- Correlation between teamTotal and Opener Batsmans
contribution
We also have correlation between Openers contribution
and the winning and losing, p-value confirm this relation
by rejecting null hypothesis



Predict match
result by using
different
features.

Regression Model with all Data

Current function value: 0.463738

Iterations 8

Intercept 0.264526

teamTotalRuns 0.120727

OpenersTotalRuns 0.015821

oppositionTotalRuns -0.128434

dtype: float64

True Positives:628

True Negatives:671

Length Test Dataset:1504

Accuracy:0.86%

The Accuracy for model with same training and test dataset is 86 %

Regression model with Splitting training and test dataset 70-30

Current function value: 0.473392

Iterations 8

Intercept 0.143369

teamTotalRuns 0.113018

OpenersTotalRuns 0.014804

oppositionTotalRuns -0.119837

dtype: float64

True Positives:196

True Negatives:201

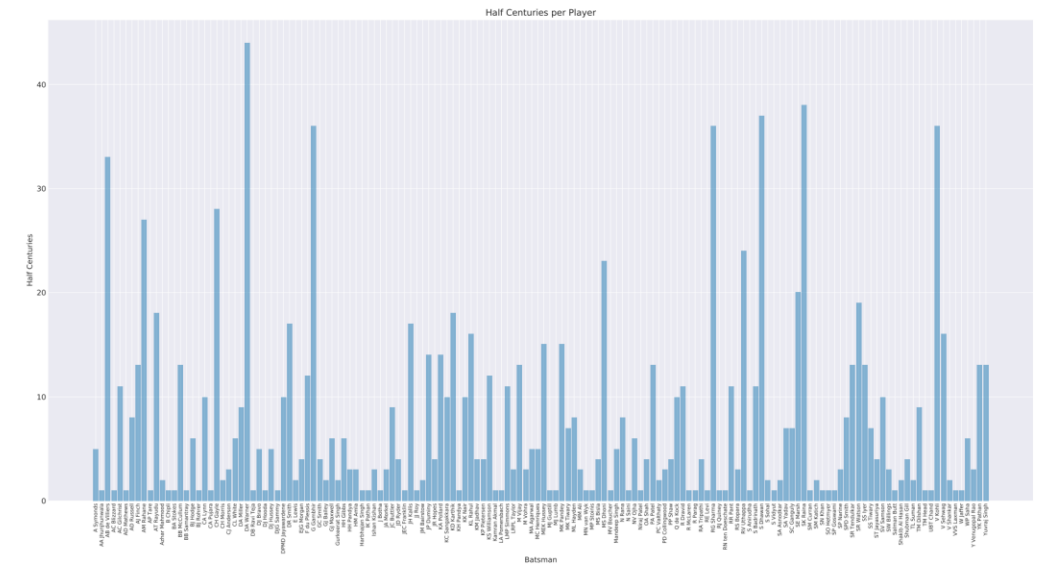
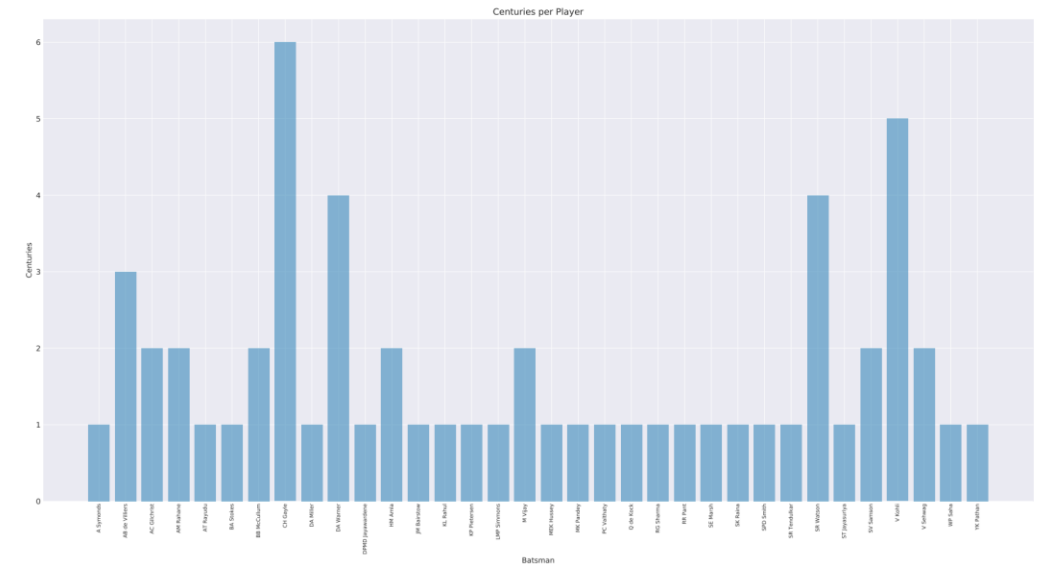
Length Test Dataset:451

Accuracy:0.88%

With splitting data into training and test dataset the accuracy is 88%

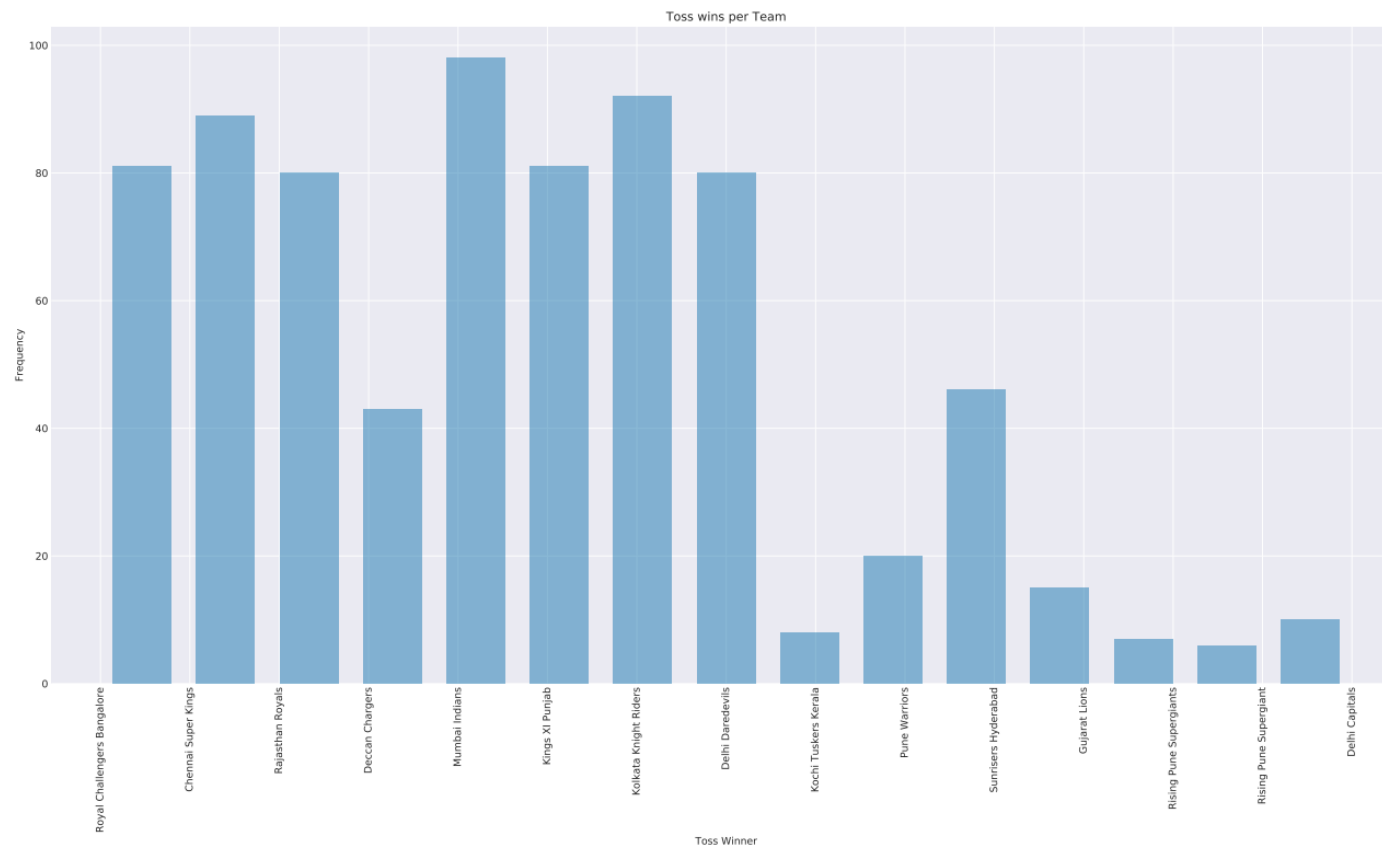
Which player got more centuries and half-centuries?

- Chris Gayle has most number centuries(6)
- David Warner has the most number of half-centuries(45)



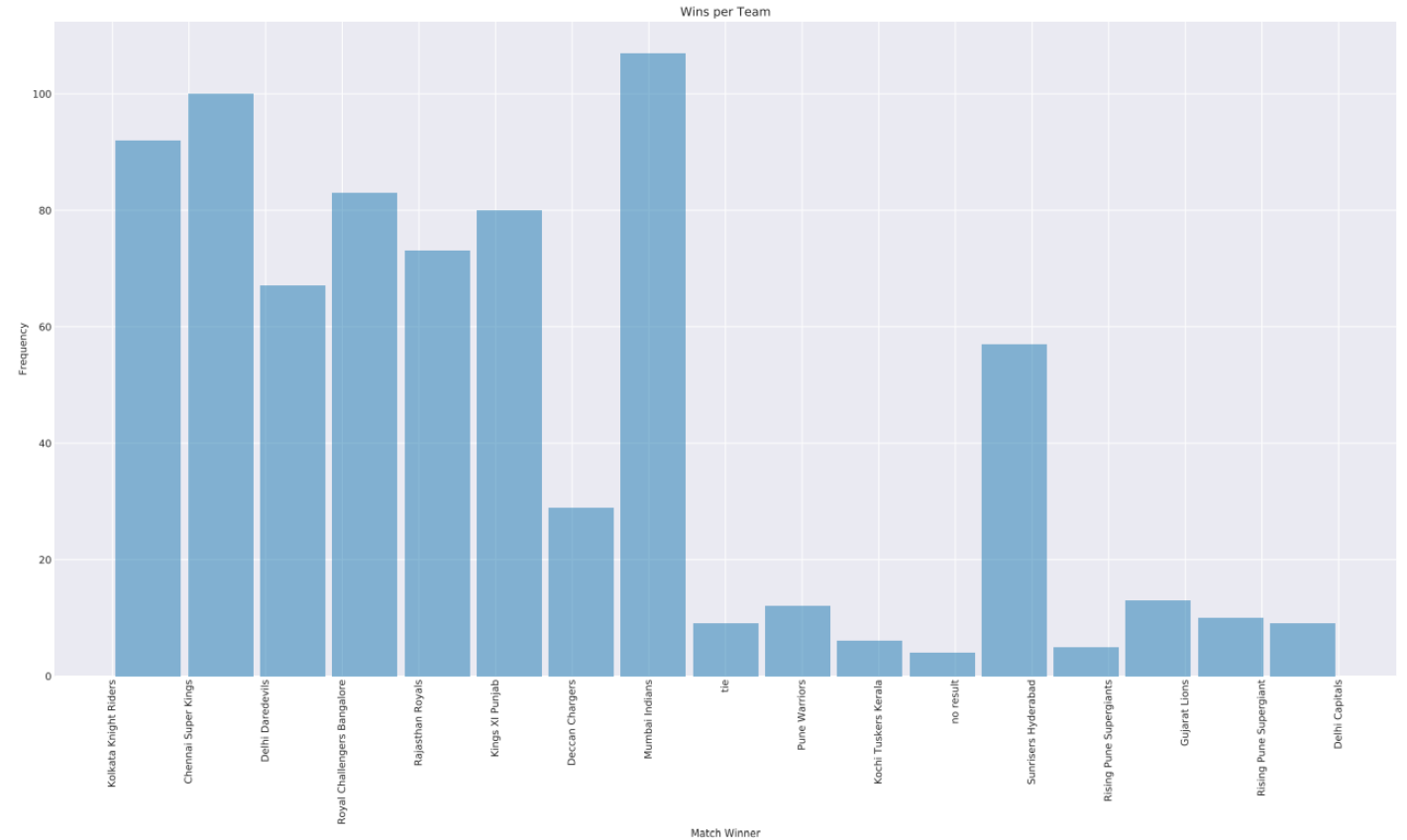
Which team won more tosses ?

- Mumbai Indians won the most number of tosses

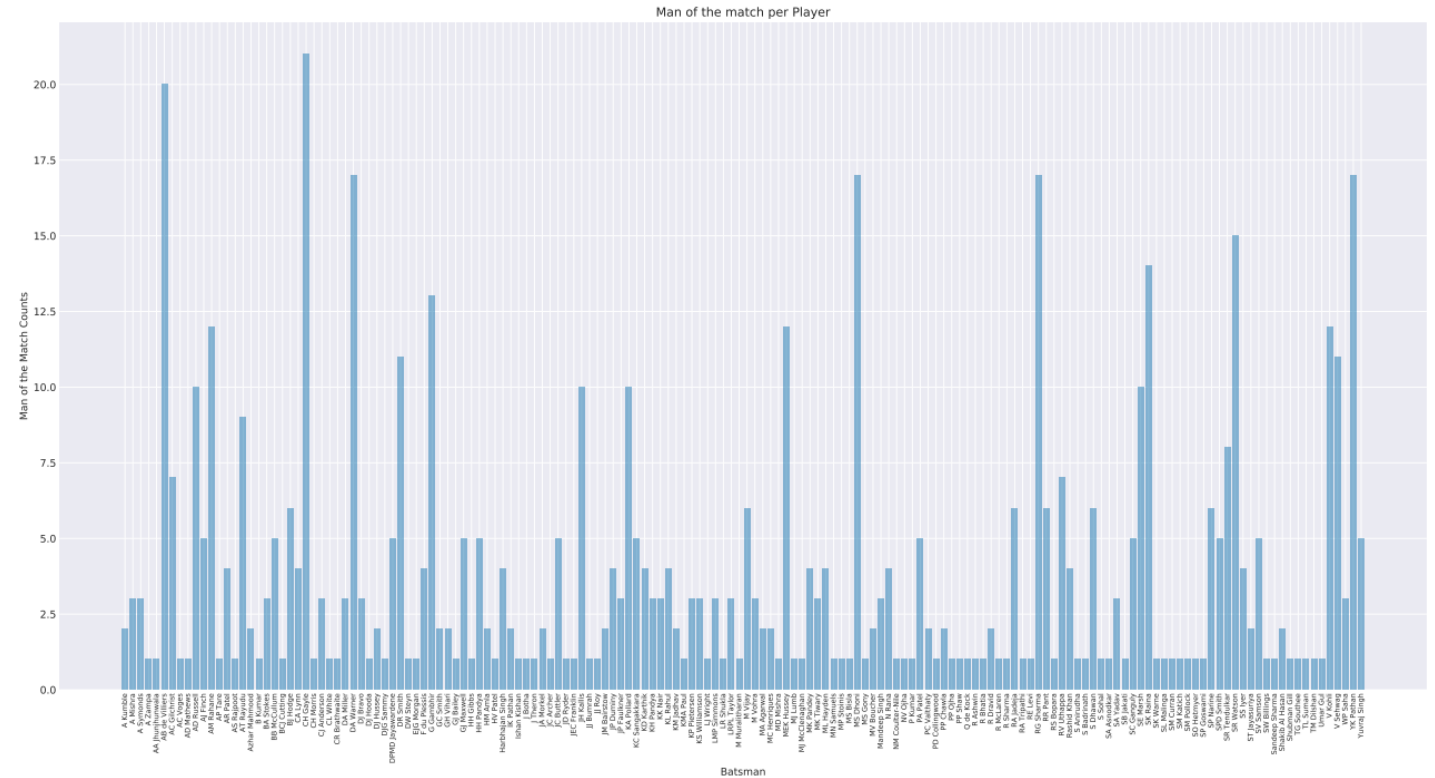


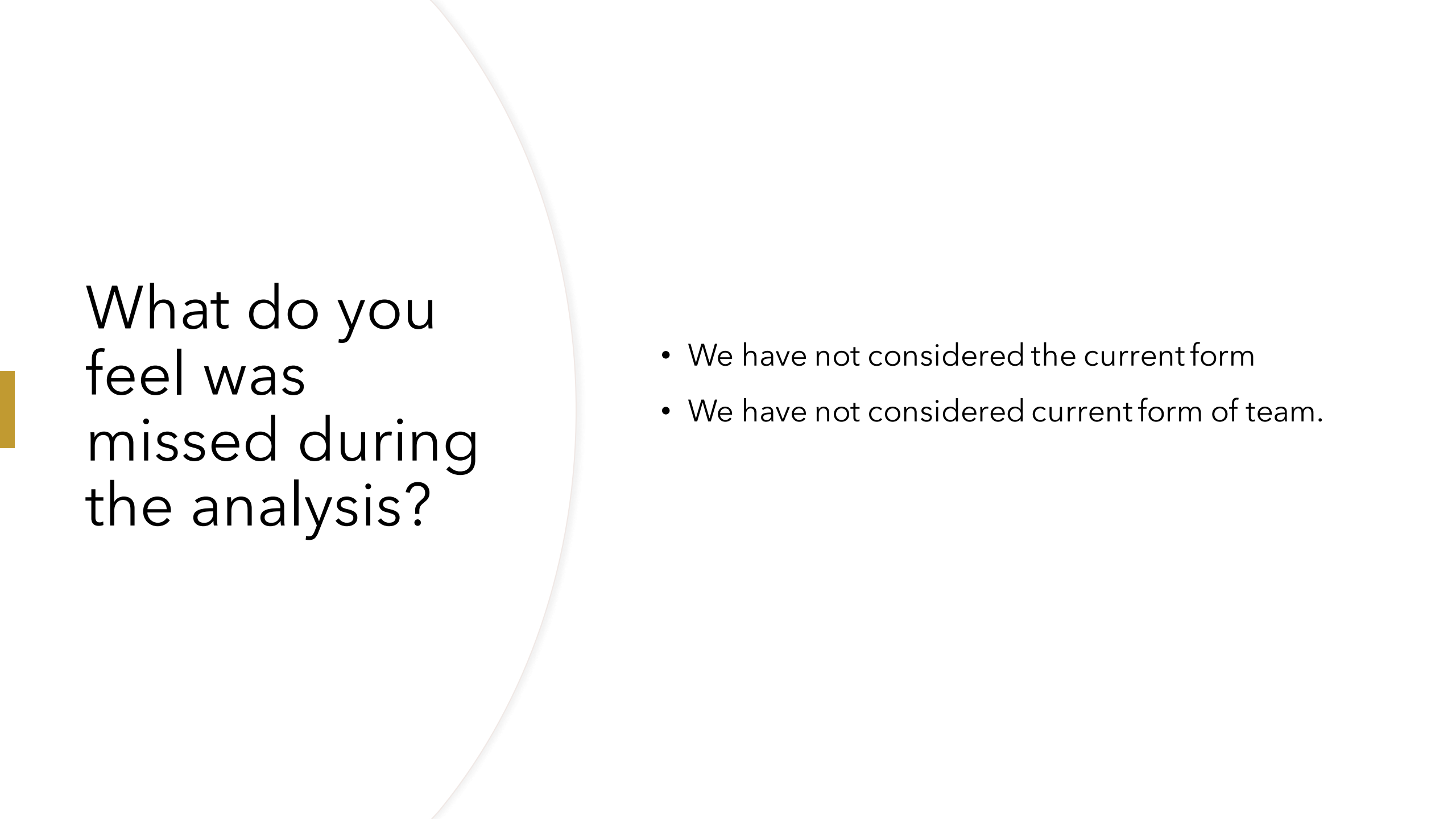
Which team won more matches ?

- Mumbai Indians won the most number of matches.




- Chris Gayle won the most number of man-of-the-match awards.





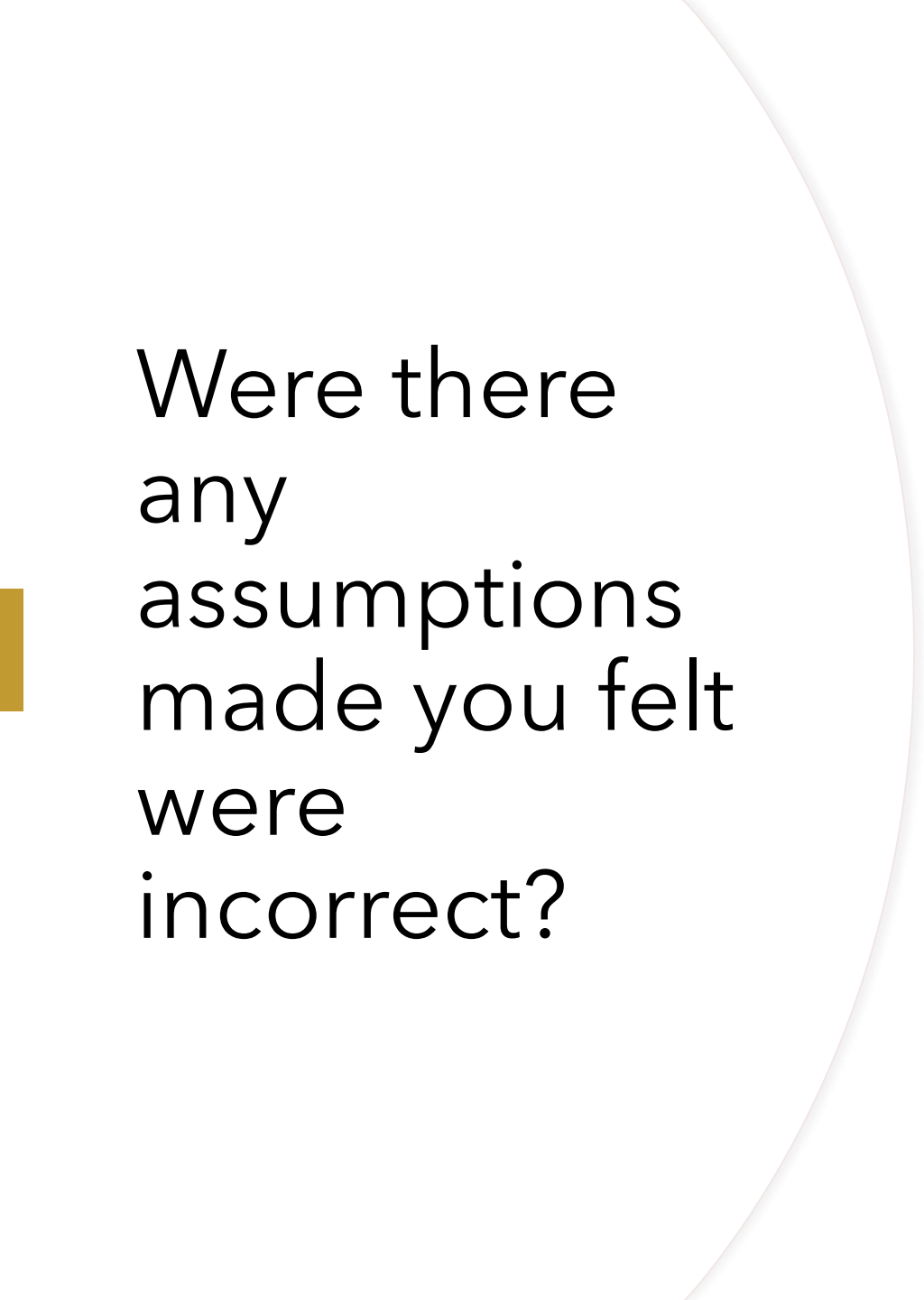
What do you
feel was
missed during
the analysis?

- We have not considered the current form
- We have not considered current form of team.




Were there
any variables
you felt could
have helped
in the
analysis?

- If data set has the wickets it would have been helped in analysis.
- Weather conditions per inning would have been helped.
- Home team variable would have helped.



Were there
any
assumptions
made you felt
were
incorrect?

- The assumption about my hypotheses questions seems to be corrected by hypothesis testing.



What
challenges did
you face, what
did you not
fully
understand?

- I have spent a lot of time in combining all the files per match into one data frame.
- I faced some challenges while adding dummy variables.

Link to Repository

- <https://github.com/sanjayjaras/-Final-Project-Data-Exploration-and-Data-Analysis>

Thank you !

