DATA ANALYSIS IN SPORTS.

INDIAN PREMIER LEAGUE

S.K. Somaiya college

Group members :

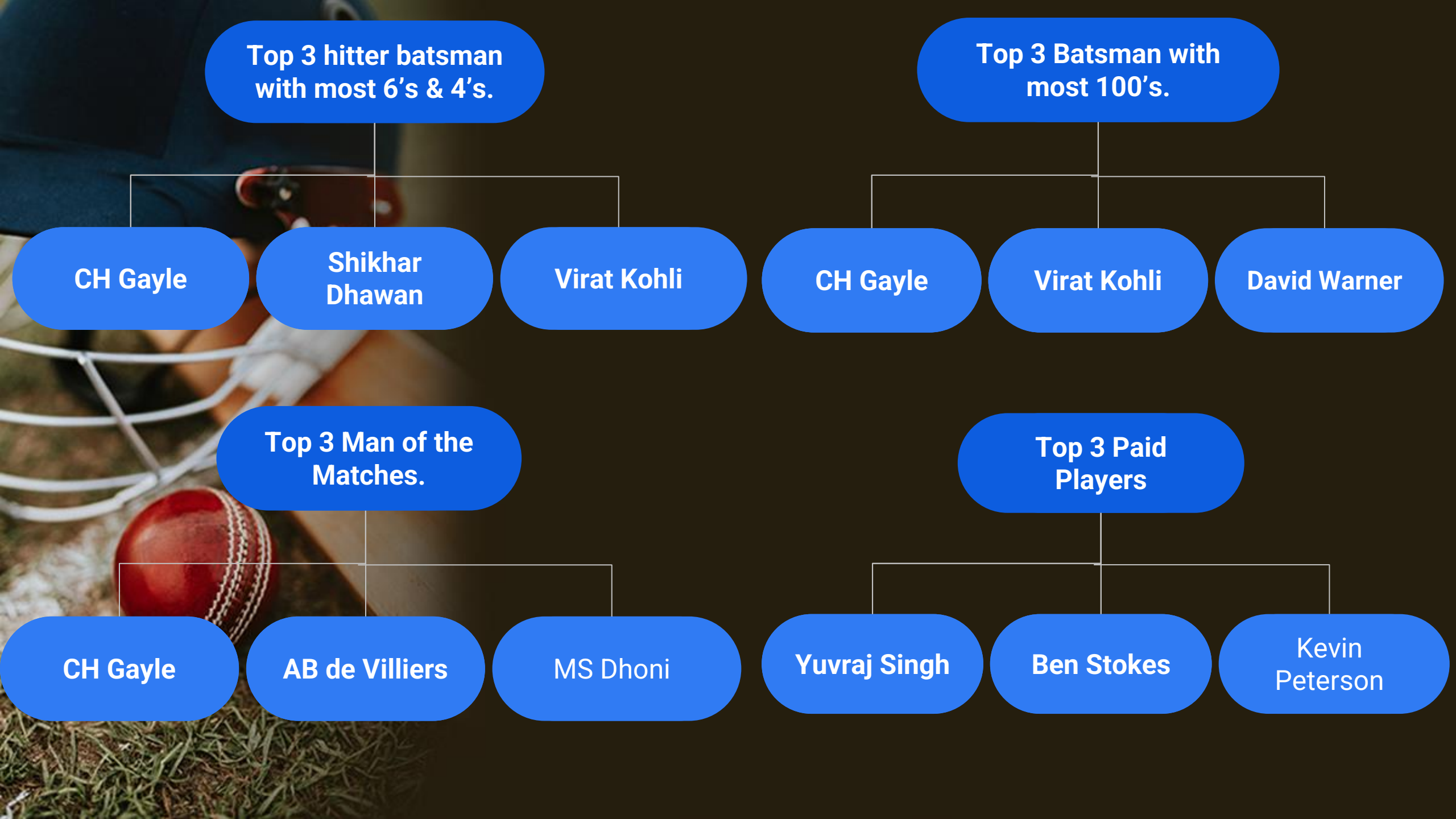
- Varsha Jadhav
- Vedang Sawant
- Tai Mohammed Moheed

Objective

- To conduct EDA and determine the best-offs of various variables.
- Comparison of average money spent by each team in each year.
- Toss and match wins, home and away wins comparison
- Prediction of position of a player in bowler and batsman standings, total runs scored by each team.

Methodology & Techniques.

- Exploratory Data Analysis.
- Analysis Of Variance.
- Multiple linear regression model and Random Forest Regression.
- R-software, Excel, Python and SPSS software were used to perform Data Analysis.
- The dataset was collected from Kaggle.



Top 3 hitter batsman with most 6’s & 4’s.

- **CH Gayle**
- **Shikhar Dhawan**
- **Virat Kohli**

Top 3 Batsman with most 100’s.

- **CH Gayle**
- **Virat Kohli**
- **David Warner**

Top 3 Man of the Matches.

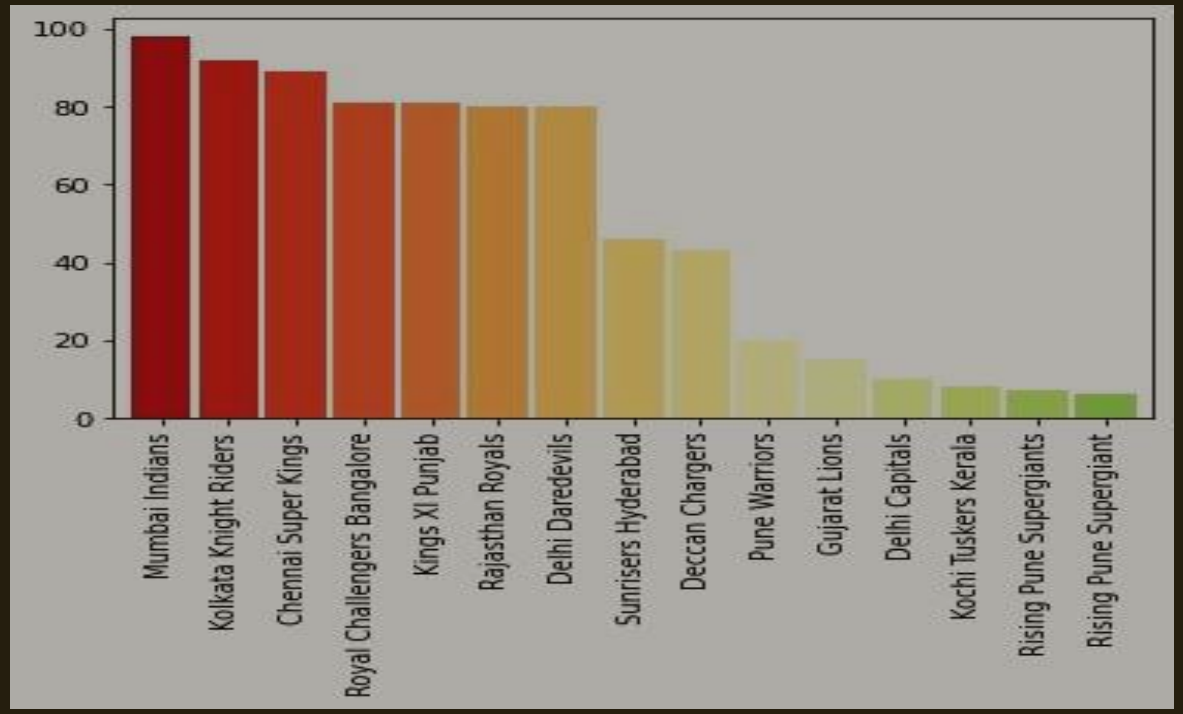
- **CH Gayle**
- **AB de Villiers**
- MS Dhoni

Top 3 Paid Players

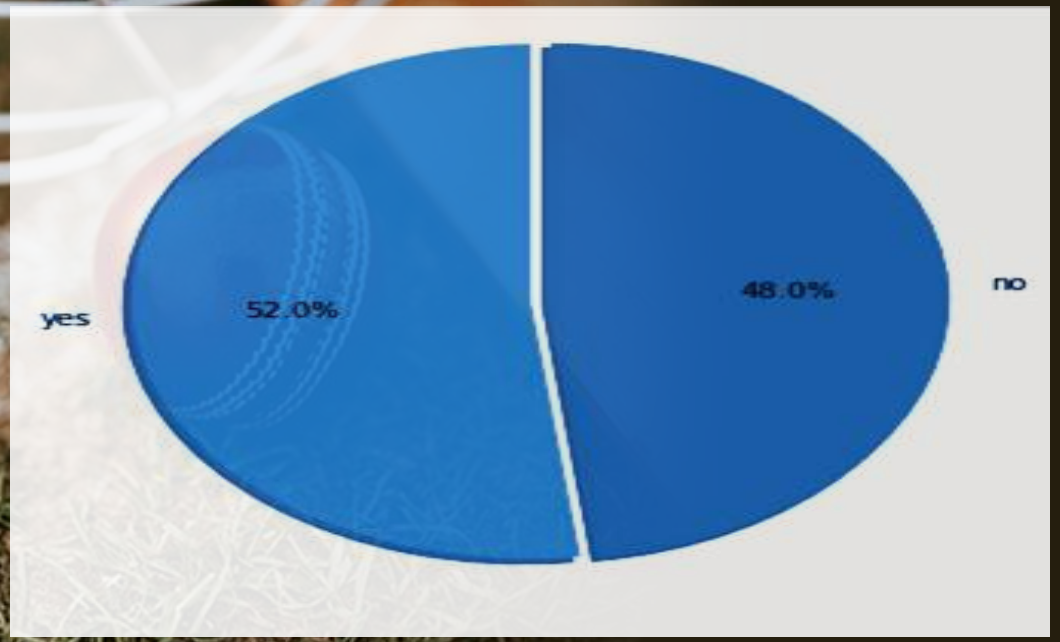
- **Yuvraj Singh**
- **Ben Stokes**
- Kevin Peterson


Top 3 Bowlers
Lasith Malinga
Amit Mishra
Piyush Chawla
Top 3 Teams.
Mumbai Indians.
Kolkata Knight Riders.
Chennai Super Kings.
Top 3 Venues
Eden Gardens
Wankhede stadium
M Chinnaswamy stadium
Top 3 Umpires.
S Ravi
HDPK Dharmasena
C Shamsuddin

Maximum Toss Winners

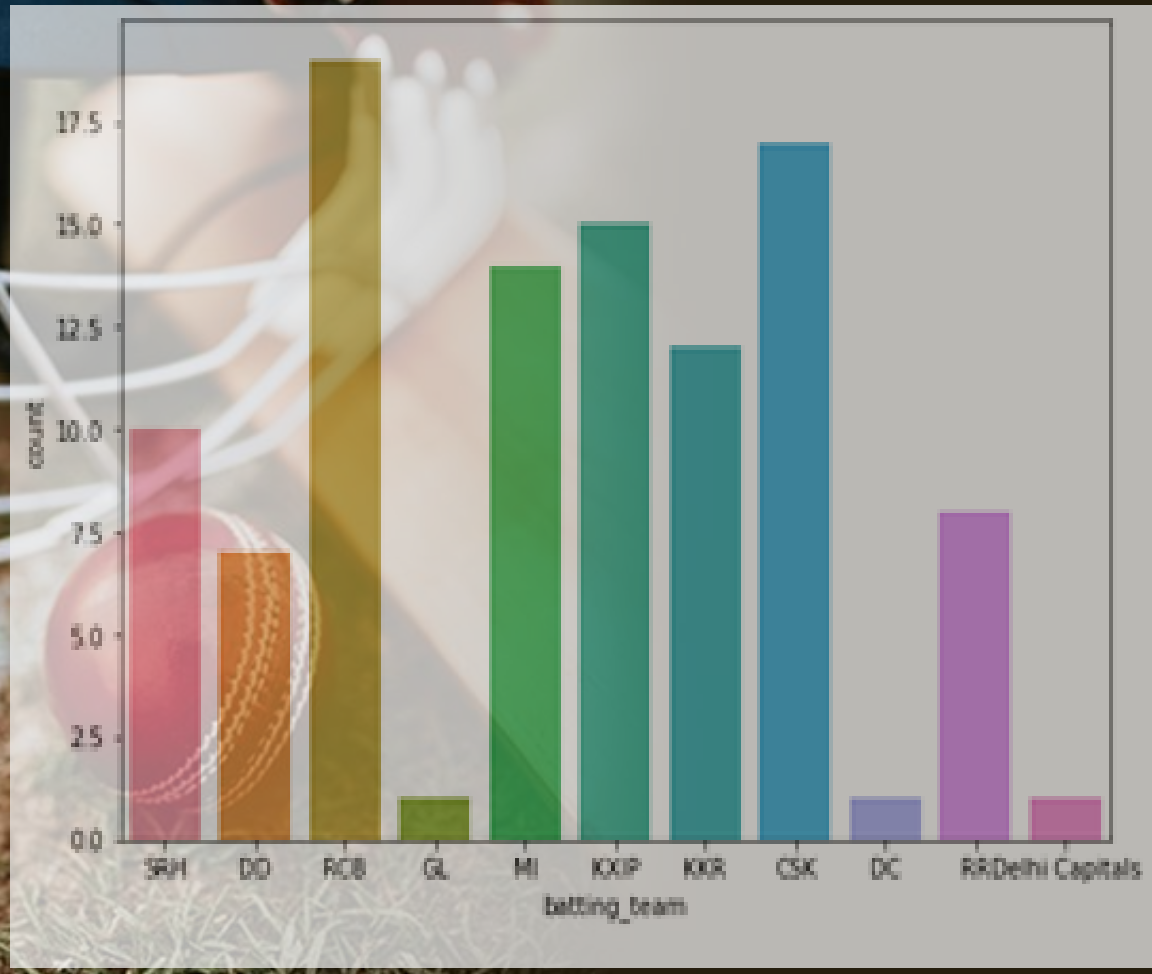


Is Toss Winner Also the Match Winner?

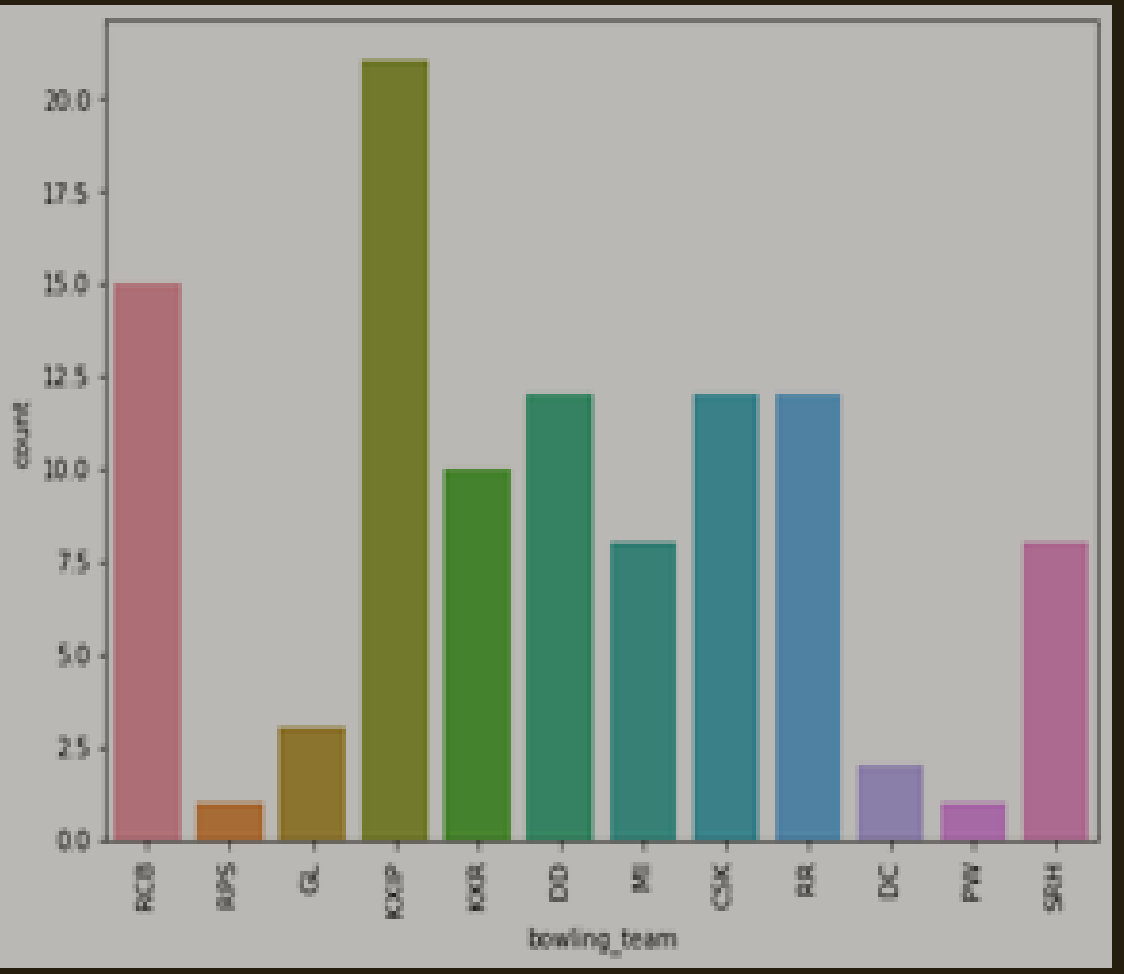


Choice after winning the toss.
Opted to Bowl 1st **61%** time.
Opted to Bat 1st **39%** time.

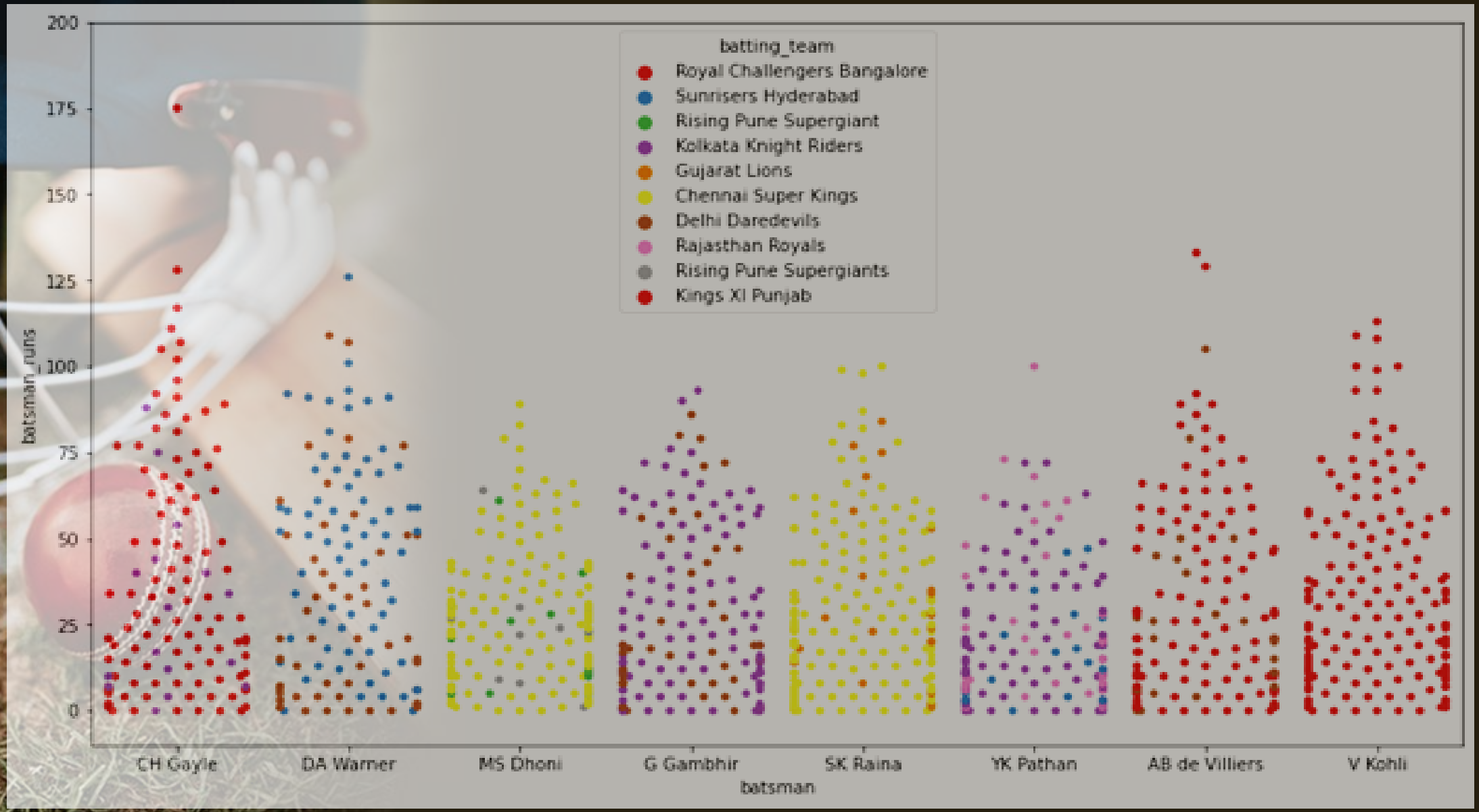
Number of times batting team has scored above 200 runs.



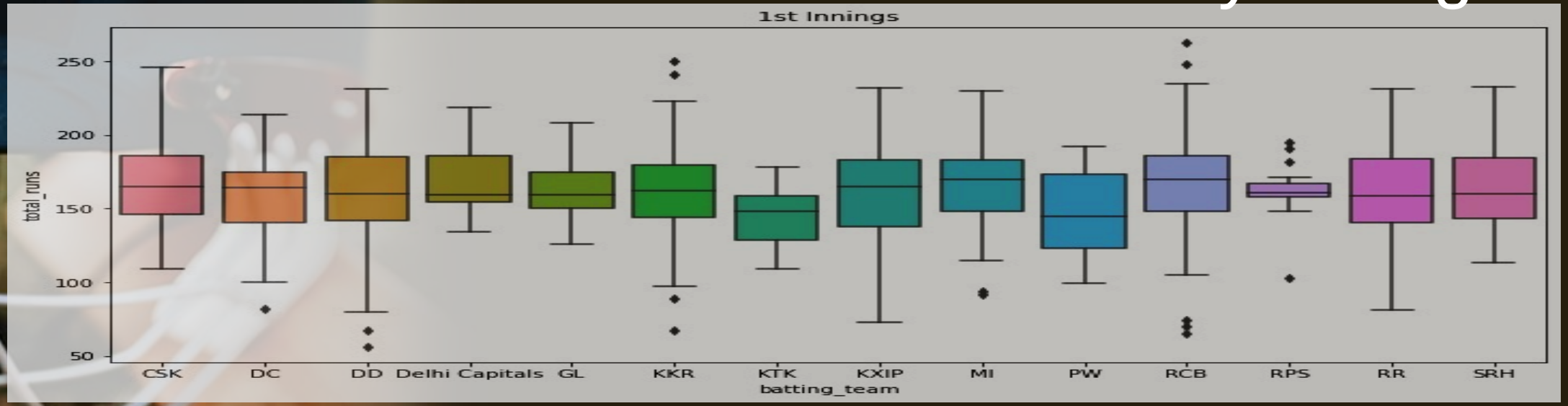
Number of times bowling team has scored above 200 runs.

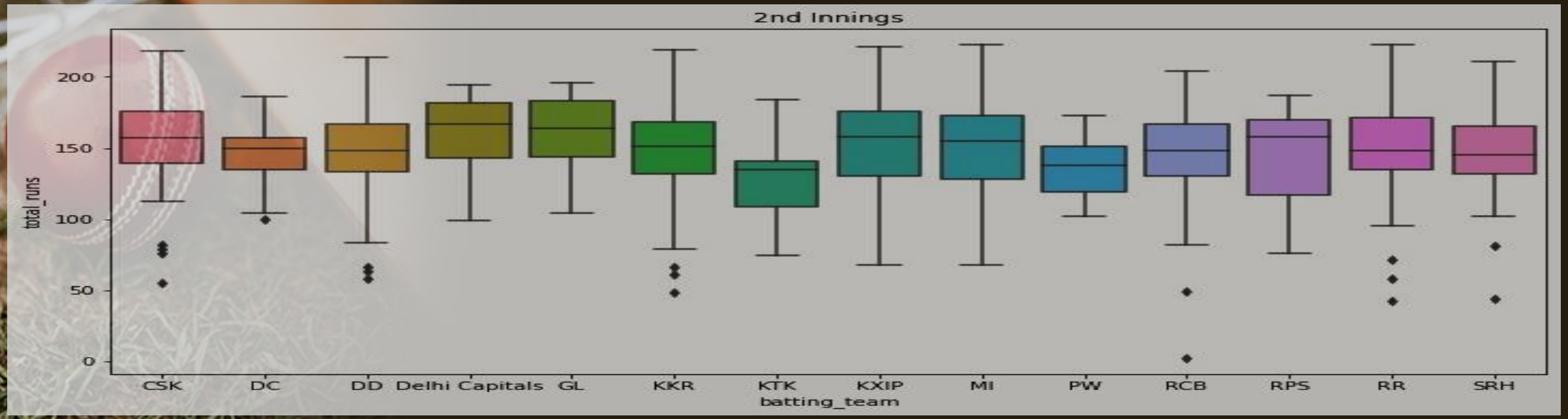


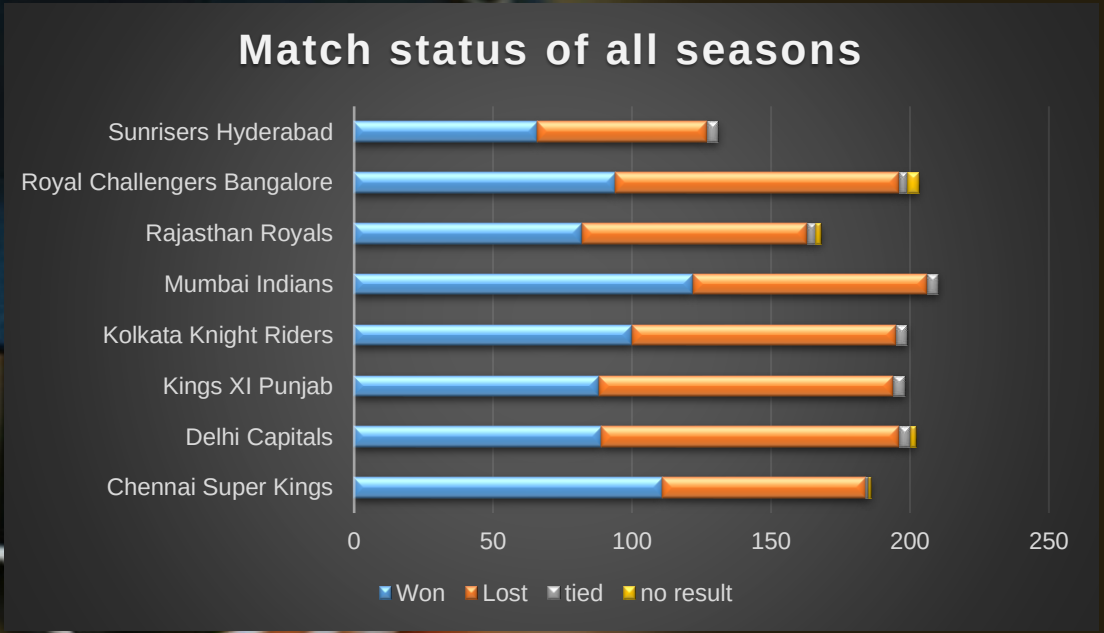
Individual Scores of Top Batsman in each Innings.

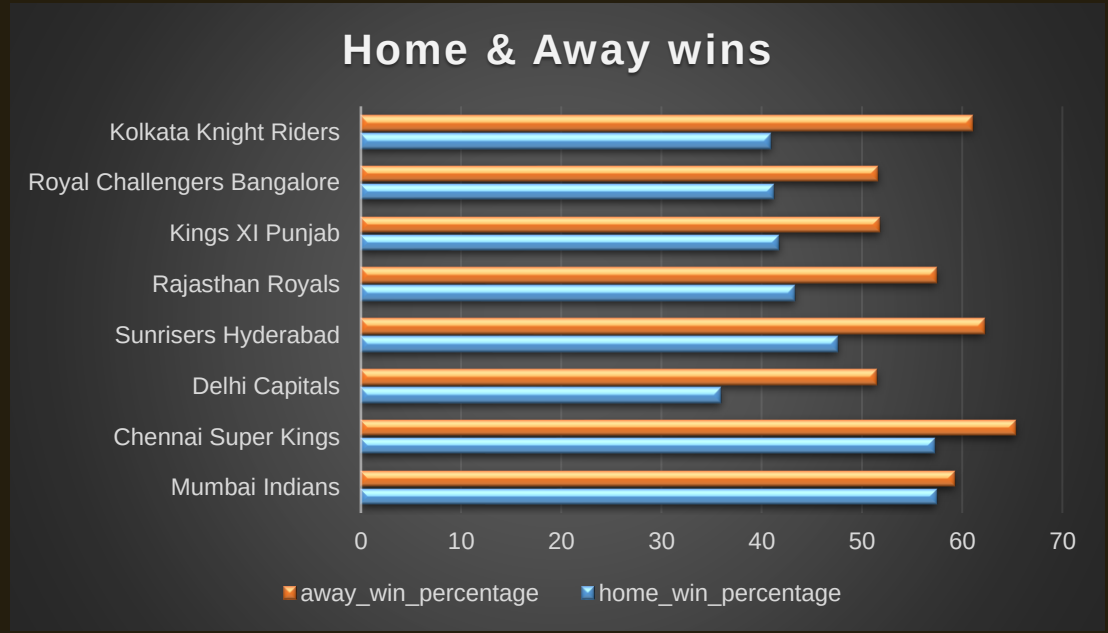


Score Distribution For Each Teams by Innings.

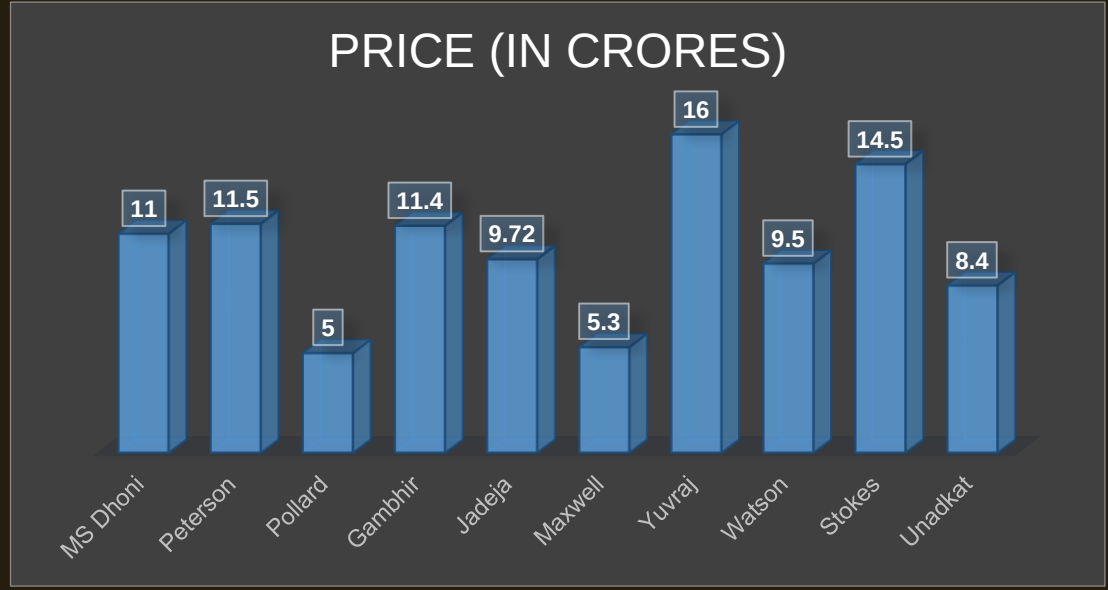





Match status of all seasons
Sunrisers Hyderabad
Royal Challengers Bangalore
Rajasthan Royals
Mumbai Indians
Kolkata Knight Riders
Kings XI Punjab
Delhi Capitals
Chennai Super Kings
0
50
100
150
200
250
Won
Lost
tied
no result

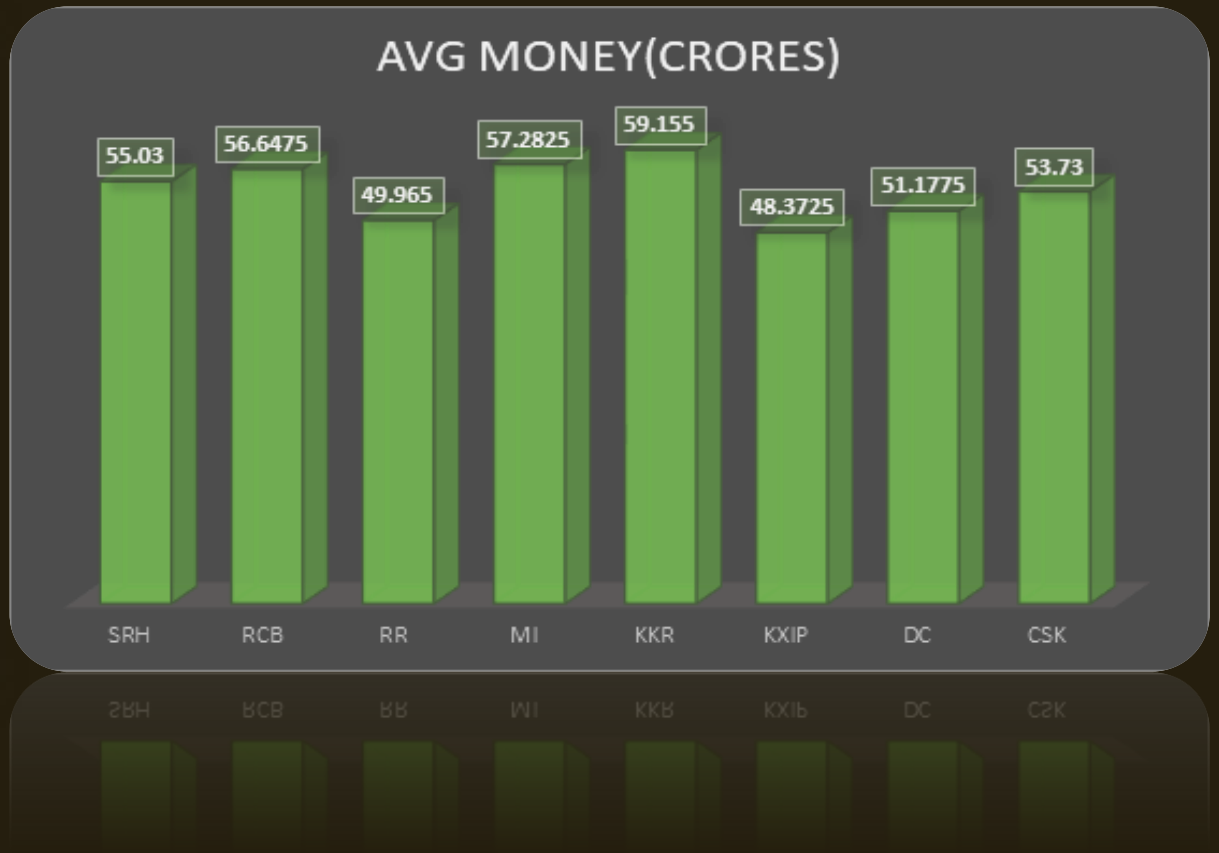

Home & Away wins
Kolkata Knight Riders
Royal Challengers Bangalore
Kings XI Punjab
Rajasthan Royals
Sunrisers Hyderabad
Delhi Capitals
Chennai Super Kings
Mumbai Indians
0
10
20
30
40
50
60
70
away_win_percentage
home_win_percentage

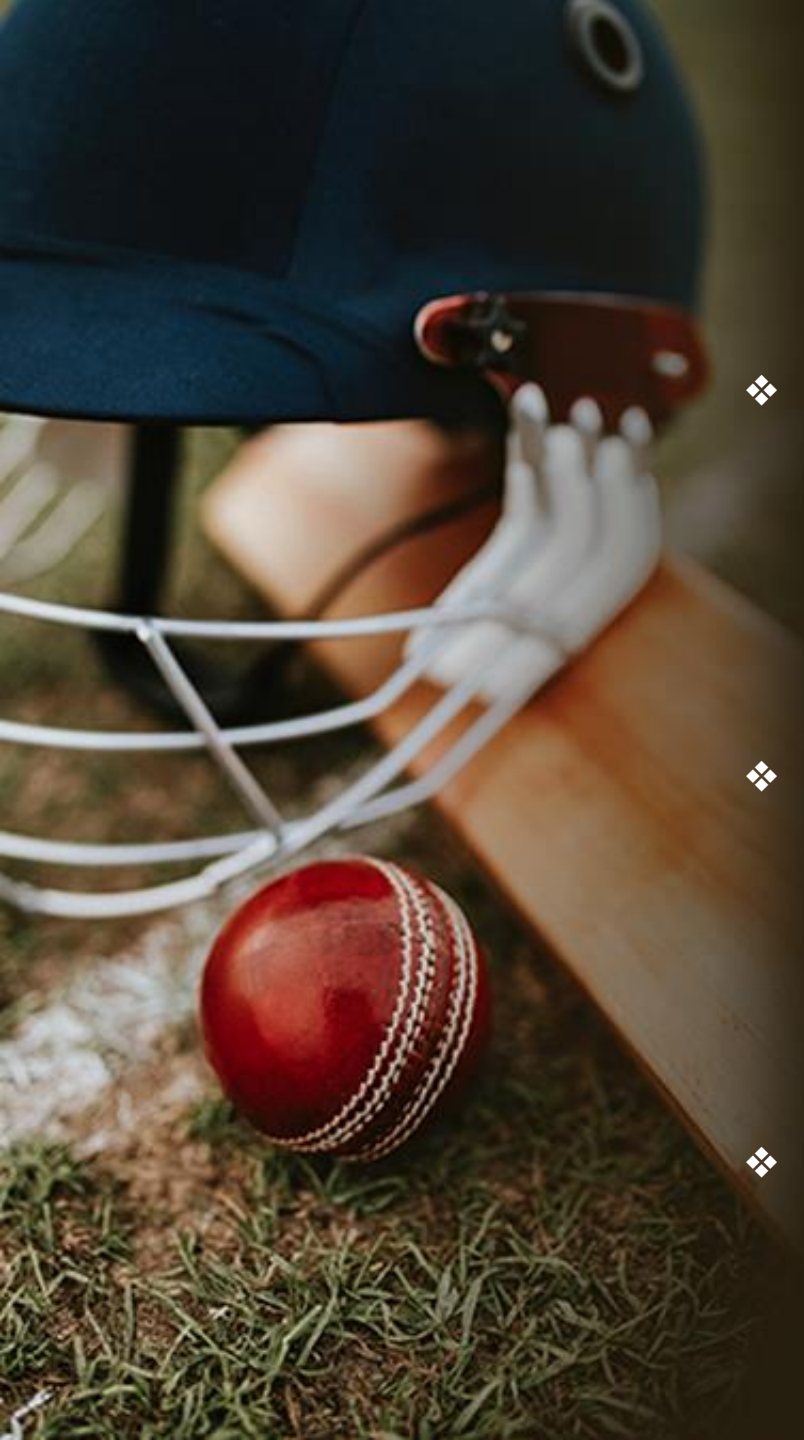

Toss & Match
60
50
40
30
20
10
0
CSK
MI
KKR
RR
RCB
KXIP
SRH
DC
toss win match win
toss win match lost


PRICE (IN CRORES)
11
11.5
5
11.4
9.72
5.3
16
9.5
14.5
8.4
MS Dhoni
Peterson
Pollard
Gambhir
Jadeja
Maxwell
Yuvraj
Watson
Stokes
Unadkat

Comparison of average money spent in each year.

```
> summary(a)
             Df Sum Sq Mean Sq F value Pr(>F)
teams         7  405.4    57.9   0.573 0.7699
year          3 1360.5   453.5   4.485 0.0139 *
Residuals    21 2123.6   101.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```





Variables under study.

- For prediction of total score at the end of the inning variables such as runs, wickets, overs, runs in last 5 overs, wickets in last 5 overs, striker & non-striker runs are considered.
- For prediction of batsman position in batsman standings variables such as matches, innings, not outs, highest score, 100s, 50s, 4s & 6s are considered.
- For prediction of bowler position in bowler standings variables such as matches, innings, overs & wickets are considered.

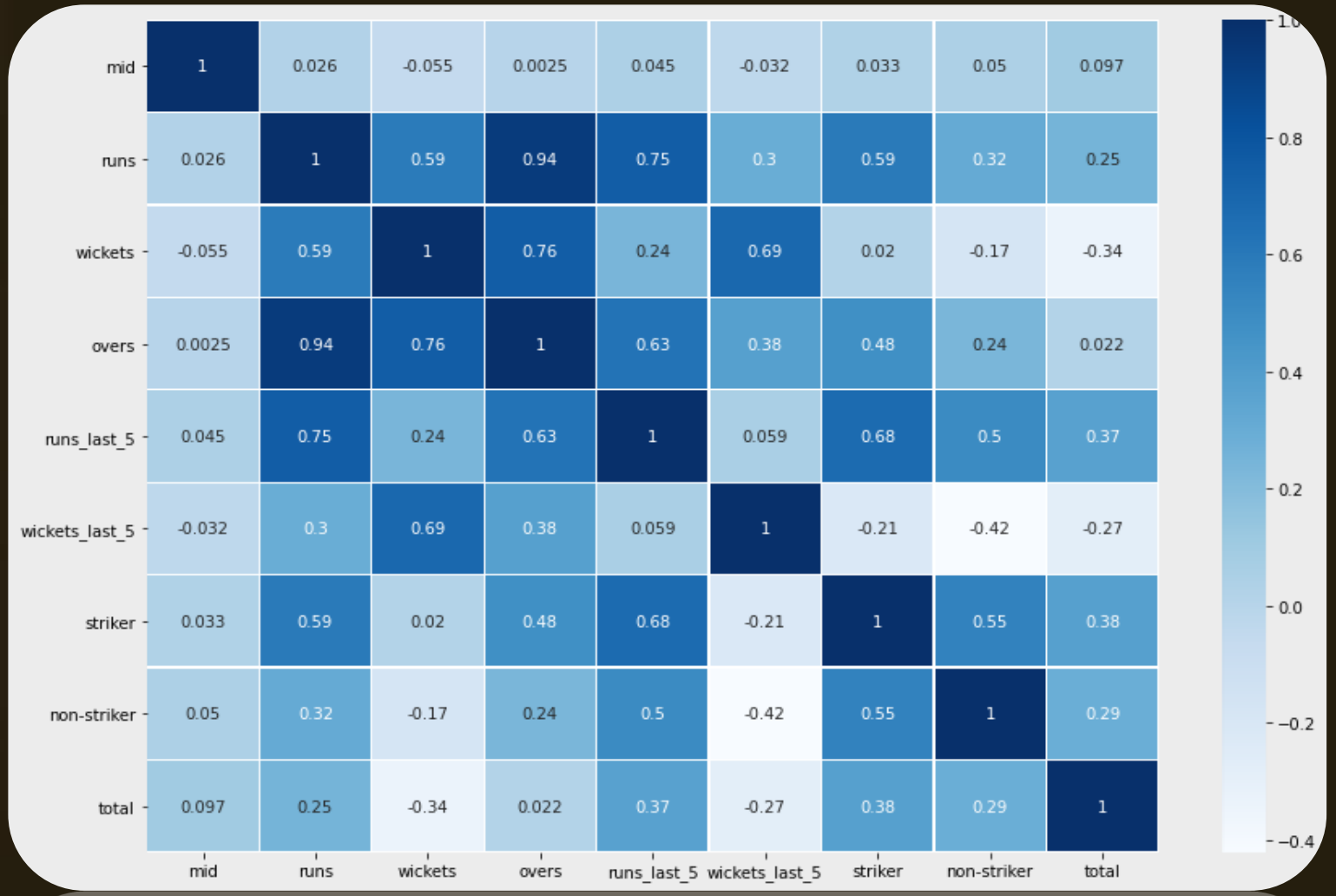
Prediction of Strike Rate.

- Variable under consideration are matches, innings, not outs, highest score, 100s, 50s, 4s & 6s.
- R-Squared value is **42.8%** when multiple linear regression was fitted to the model.
- R-Squared value is **87.89%** when Random Forest Regression was fitted to predict the strike rate.
- Random Forest regression proved to be a better fit than multiple linear regression model.

Prediction of position of batsman & bowler in standings

- Variables under consideration for prediction of batsman position in standings are matches, innings, not outs, highest score, 100s, 50s, 4s & 6s.
- R square value is **70.01%** when multiple regression was fitted to the model.
- Variables under consideration for prediction of bowler position in standings are matches, innings, overs & wickets.
- R square value is **87.47%** when multiple linear regression was fitted to the model.

Correlation Plot Between Variables



Prediction of Total Runs for Each Inning.

- Variables under consideration to predict total runs are runs, wickets, overs, runs in last 5 overs, wickets in last 5 overs, striker & non-striker runs.
- R-Squared value is **50.88%** when multiple linear regression was fitted to the model.
- R-Squared value is **67.46%** when Random Forest Regression was fitted to predict the strike rate.
- Random Forest regression proved to be a better fit than multiple linear regression model to predict the total runs for each inning.

Limitations

- Data which was considered in analysis is from 2008-19.
- Lack of information on pitch & weather conditions in matches.
- Some teams were not participated in all seasons from 2008-19.
- Data structure was not appropriate as it can large number of empty cells that was replaced by zero initially.
- Complete record of money spent by each team in all seasons was unavailable.

Observations

- **CH Gayle** is the best player with maximum runs (5 centuries), maximum boundaries and maximum man of the match title but than too he is not the most expensive player.
- **MS Dhoni** and **Gautam Gambhir** have never scored a Century.
- **V Kohli** has played only for Royal Challengers Banglore in all seasons.
- **Mumbai Indian** is the best team and **Eden Gardens** is the best venue and **S Ravi** is the favorite umpire.
- **Yuvraj Singh** is the highly paid cricketer and **Lasith Malinga** is the bowler with maximum wicket.
- **Mumbai Indians** seem to be very lucky having the highest win in tosses followed by **Kolkata Knight Riders**. **Pune Supergiant s** have the lowest wins as they have played the lowest matches also. This does not show the higher chances of winning the toss as the number of matches played by each team is uneven.
- **RCB** is the team which cross 200 runs while setting up a target maximum number of times followed by **CSK**, **KXIP** and **MI**. While chasing the target **KXIP** cross the target of 200 maximum times followed by **RCB**.
- According to Box & Whisker plot the batting by **CSK** in innings 1 looks to be the best. Innings 2 also conveys the same story.
- **KKR** spends the maximum money on teams followed by **MI** and KXIP spend least money.
- Correlation between total runs and the striker batsman runs is maximum with 0.38, followed by run last 5 overs with 0.37 and non-striker batsman runs with 0.29 and these validates our data is true.

Interpretation

- The toss winner is not necessarily the match winner. The match winning probability for toss winning team is about 50%-50% almost same.
- MI & CSK won maximum matches in all seasons of IPL. RCB, DC & KXIP lost maximum matches as compared to wins. Maximum number of tie had occurred in RCB & DC matches. Maximum no results were found in RCB matches.
- Match wins of MI & CSK are more as compared to toss wins. Match wins of DC & KXIP are less as compared to toss wins. SRH & RCB had approximately equal amounts of match & toss wins.
- KKR, CSK & SRH won more away matches as compared home matches. MI won approximately equal home & away matches.
- Average Money Spent during each season of IPL is same but money spent on each team differs and this is validated by Two-Way ANOVA Model.
- Random Forest Regression is the best fit to predict the total runs with 67.46% accuracy, also Random Forest Regression seems to be the best fit to predict strike rate with 87.89% accuracy.
- Multiple linear regression is the best fit to predict both bowler & batsman position in standings with 87.47% and 70.01% accuracy respectively.

Thank you !