Indian Premier League
Season 2008-2019

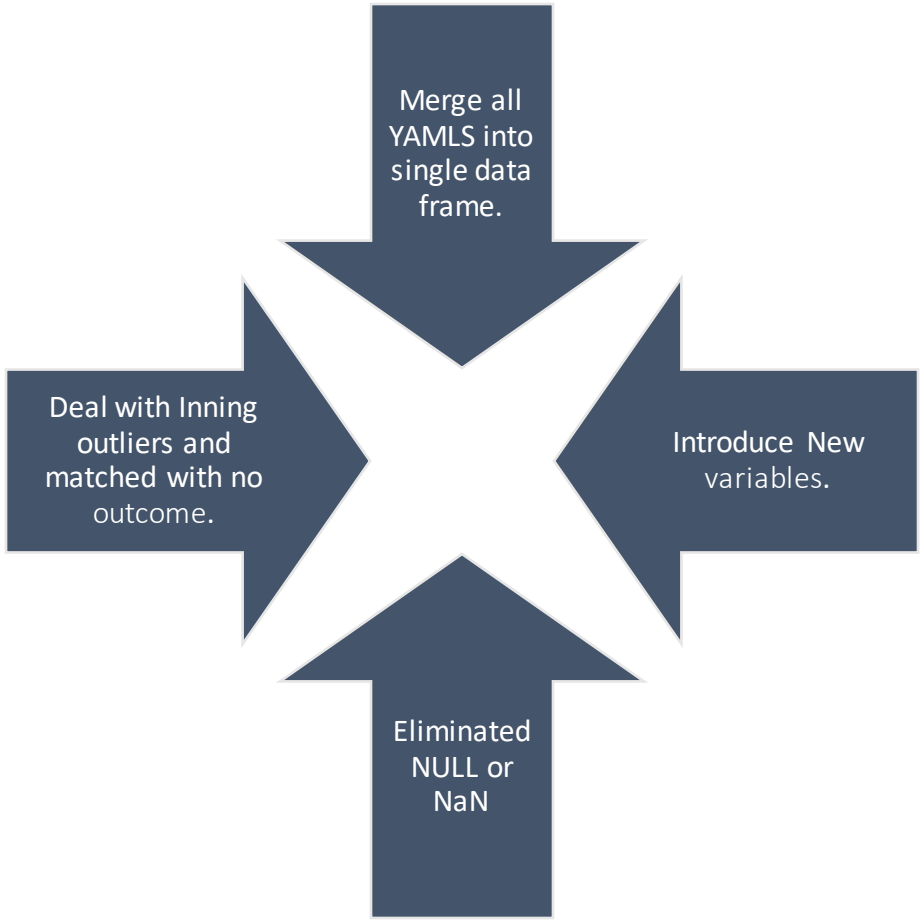
The Indian Premier League (IPL) is a professional Twenty20 cricket league in India. It was founded by the Board of Control for Cricket in India (BCCI) in 2007 and is currently the most attended cricket league in the world. The IPL features teams representing various cities and states in India

Data Source : Indian Premier League 2008-2019 | Kaggle

Questions – Statistical and Hypothetical

- Average innings scores
- Chennai Supper Kings – Really a king ?
- Openers Contribution.
- Best batter
- Predict the match result
- Toss analysis
- Man of the Match
- First and Last six overs impact.

Data Preparation and Integration



Average Team Scores Over the Seasons

Team average score over is season is increasing ?

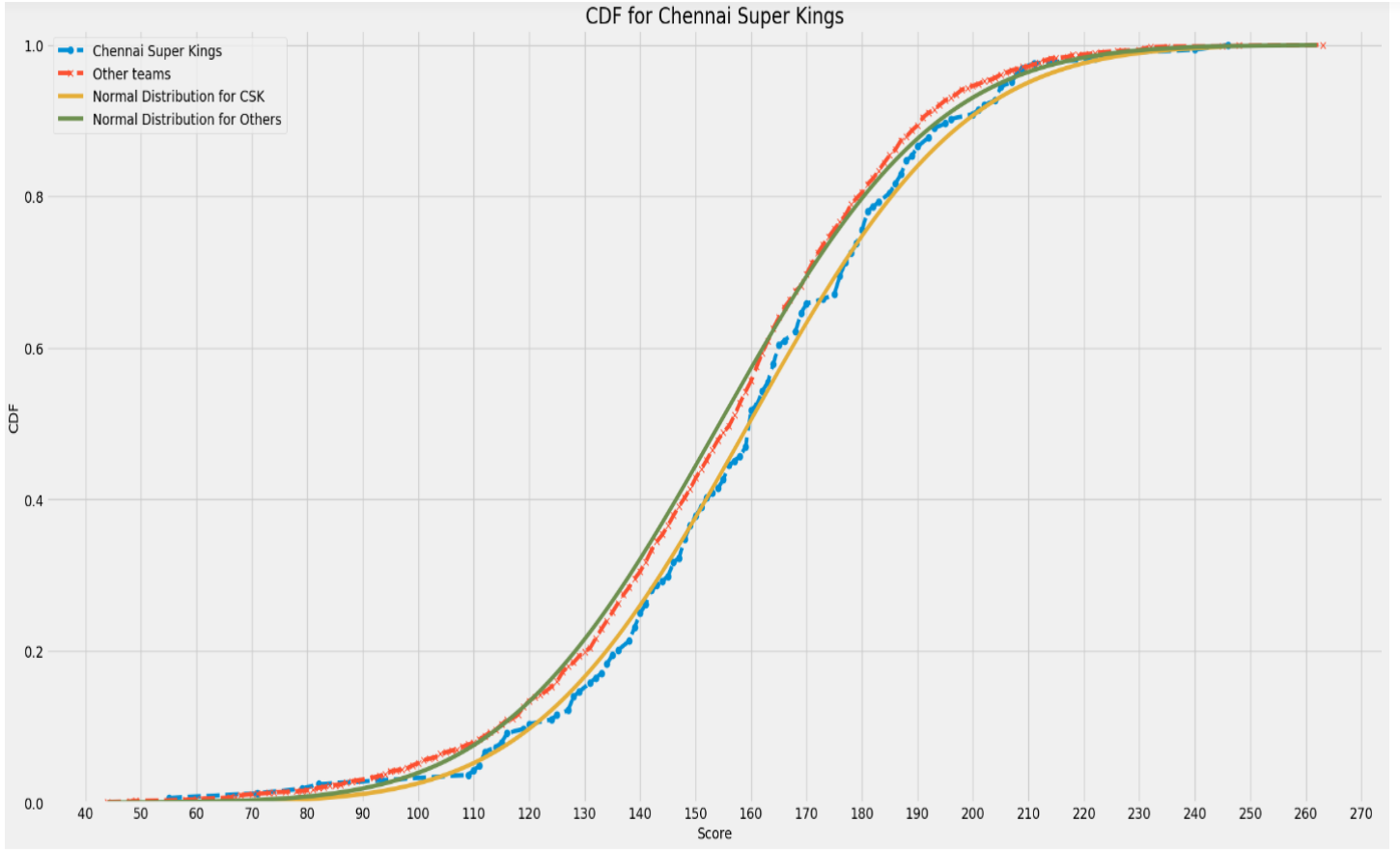
Scatter plot with mean() by season indicates average increase in score with the seasons.

IPL Seasons:

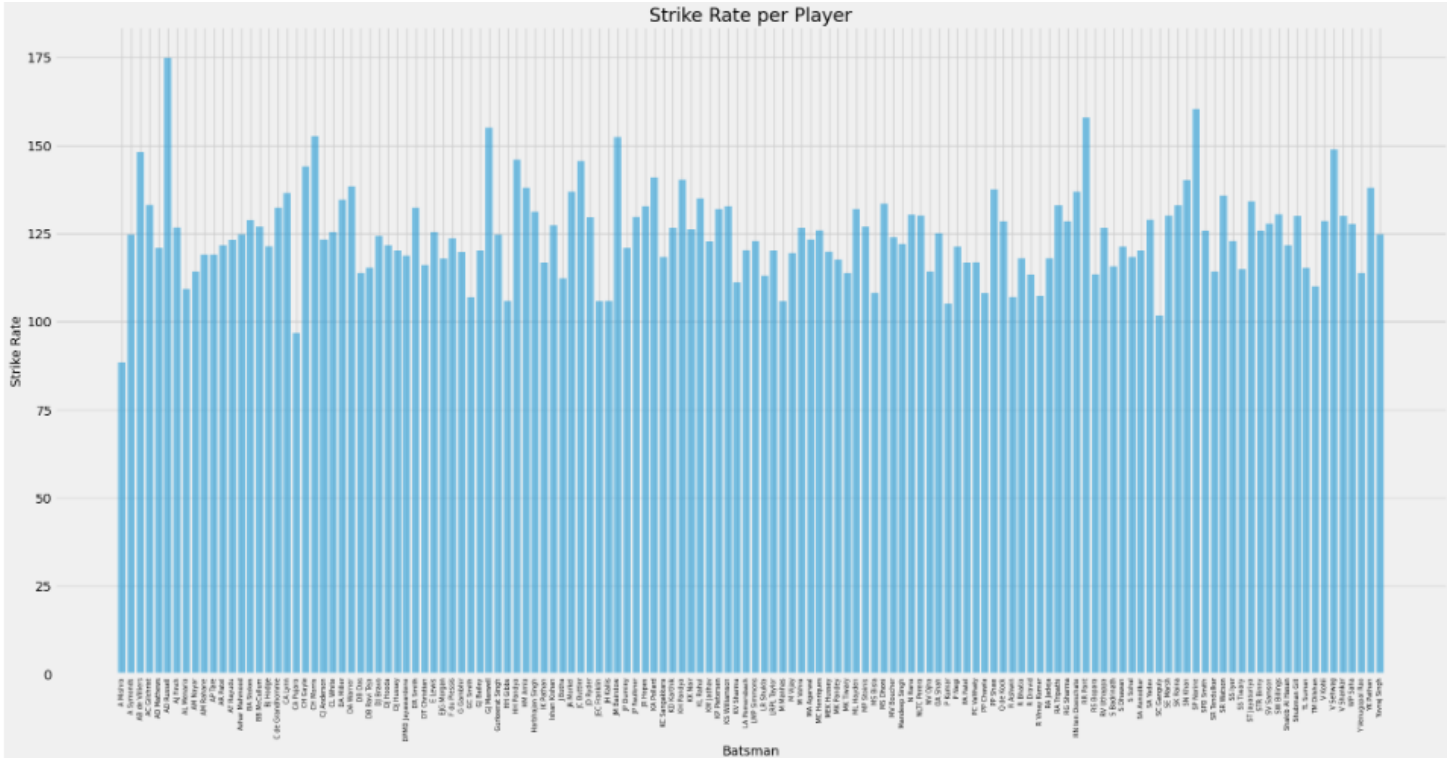
2008	154.629310
2009	143.157895
2010	157.200000
2011	146.513889
2012	151.709459
2013	148.296053
2014	157.575000
2015	157.394737
2016	157.183333
2017	159.059322
2018	165.841667
2019	163.533898

Data Distribution-CDF

- Data is normally distributed.
- Normal Distribution for CSK Mean:159.08 Var:933.08 Sigma:30.55
- Normal Distribution for Others Mean:154.27 Var:953.29 Sigma:30.88



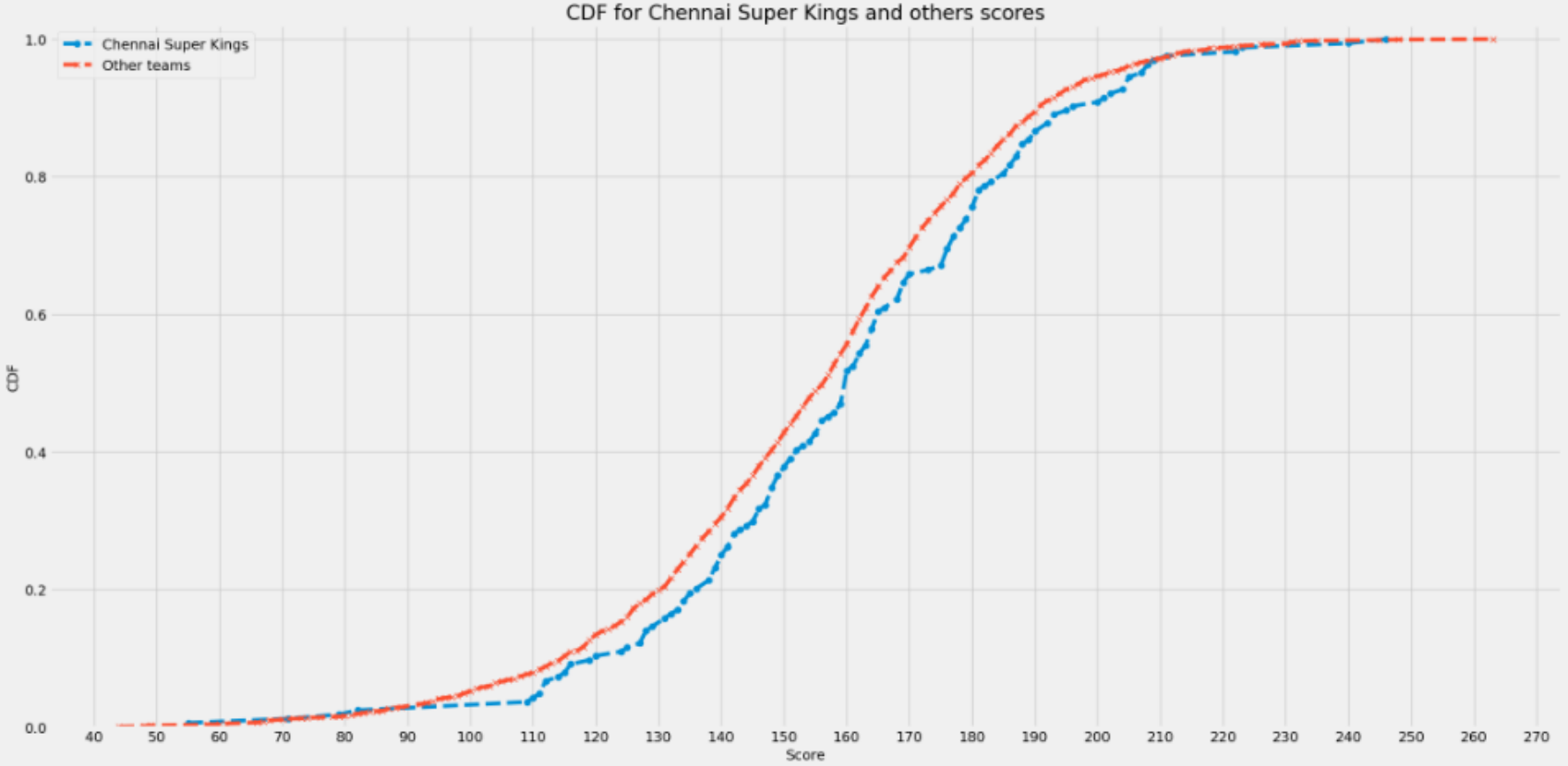
Batsman with strike rate

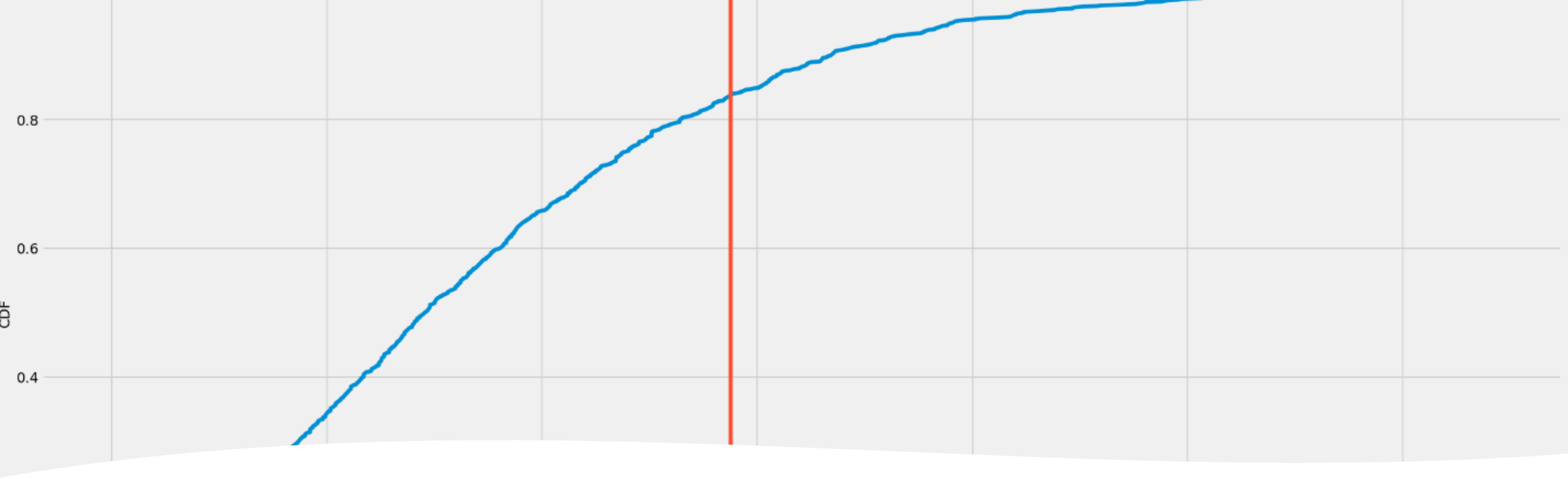


- Batsman with best strike rate?
- Aundre Russel with strike rate 175, is the best from plotted Histogram

Chennai Super Kings is really a King ?

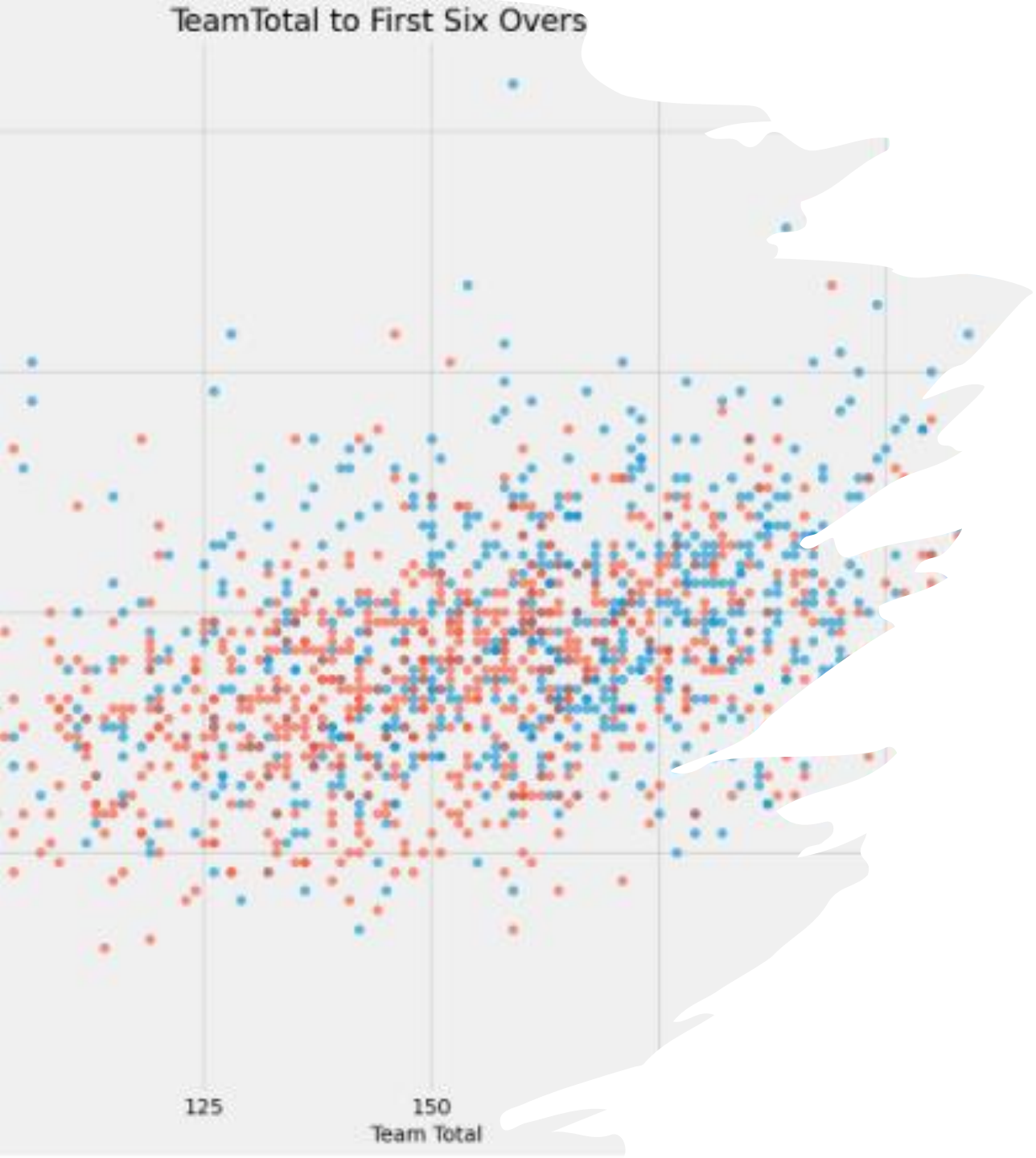
The comparison CDF shows CSK did scores better between 170 to 200, but average for all other teams are better than CSK.





King Kholi Second Inning Analysis – Virat Kholi

- Scores comparison using hypothesis test for Virat Kohli.
- The PValue (i.e 0.16) indicates 2nd innings score as big as the observed difference 5.76, i.e 16% of the first inning score.
- End result is not statistically significant, however CDF intersects at 0.84 complement of the PValue 0.16.



Impact of first 6 overs on match result?

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

SignificanceResult(statistic=0.2283302908772324, pvalue=3.066031954648354e-19)

Above correction 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.

Impact of last 5 overs on match result?

Covariance

	teamTotalRuns	lastFiveTotal
teamTotalRuns	954.461770	418.696843
lastFiveTotal	418.696843	348.678816

Pearson Correlation

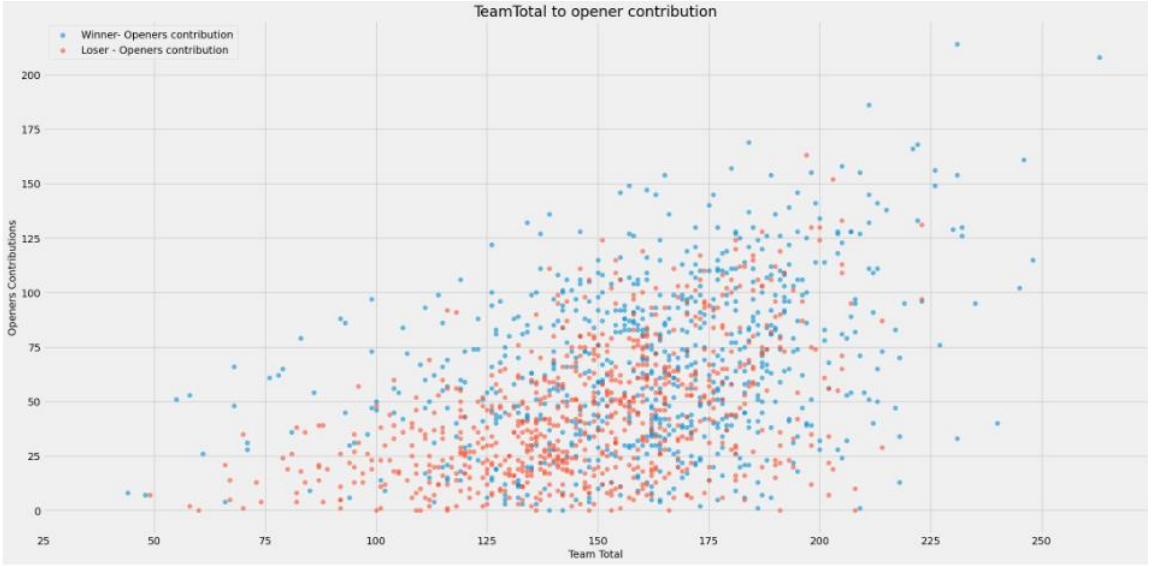
	teamTotalRuns	lastFiveTotal
teamTotalRuns	1.000000	0.725784
lastFiveTotal	0.725784	1.000000

SignificanceResult(statistic=0.026314742201529143, pvalue=0.30779881584986263)

- Matrix result confirms positive relationship between last five overs score and total team score.
- Point biserial Correlation values shows positive relationship however p-value is not rejecting null hypothesis as p-value is greater than 0.05.

Opener batting contribution influence match result ?

Correlation matrix indicates correlation between Openers contribution and end results (winning and losing). p-value confirm this relation by rejecting null hypothesis



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

SignificanceResult(statistic=0.3064424433593572, pvalue=4.583588404216702e-34)

Predict match outcome :

Regression Model - Accuracy for model entire data set is 86 %

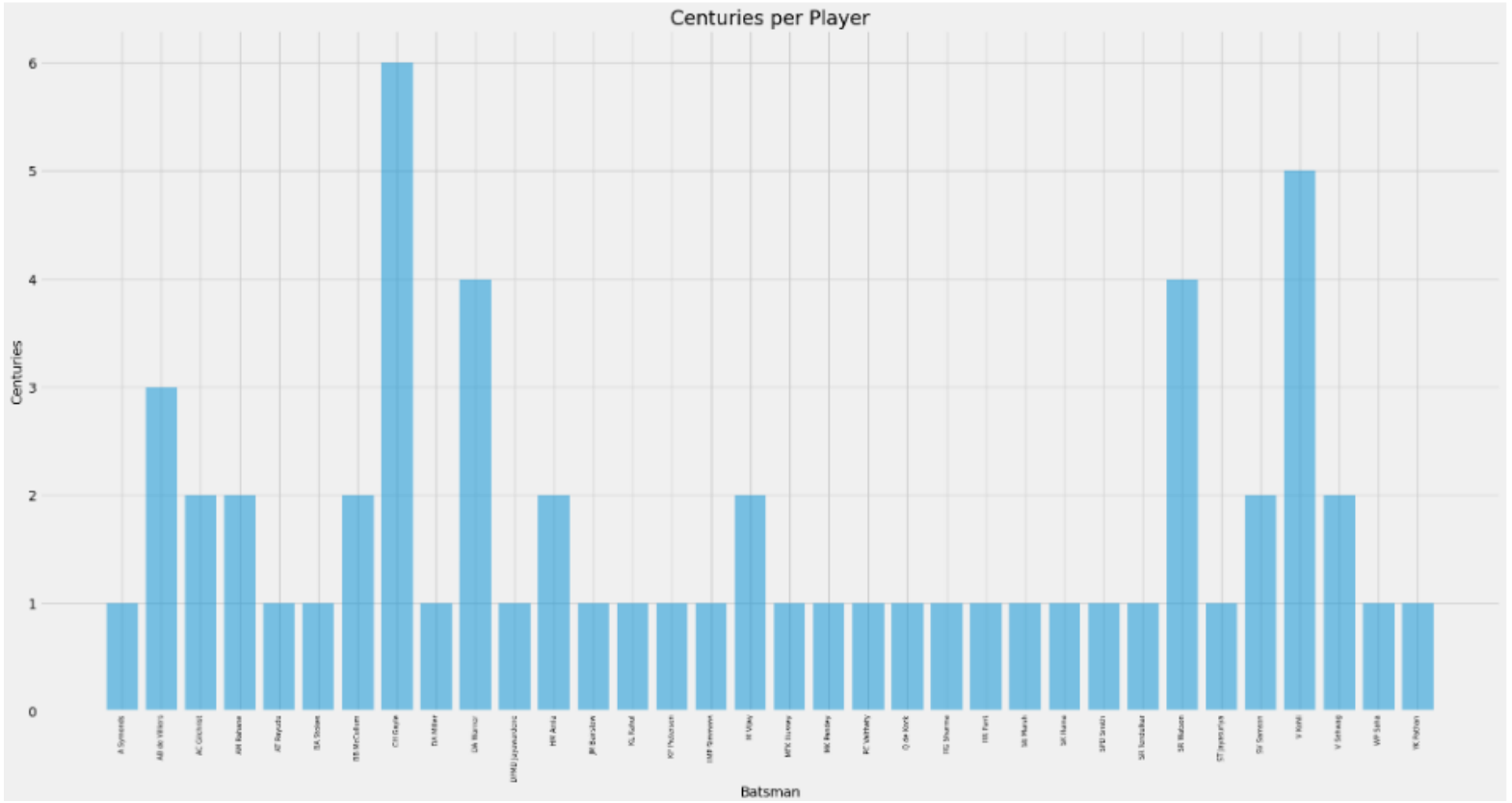
```
Optimization terminated successfully.
         Current function value: 0.463738
         Iterations 8
Intercept              0.264526
teamTotalRuns          0.120727
OpenersTotalRuns       0.015821
oppisitionTotalRuns   -0.128434
dtype: float64
True Positives:628       True Negatives:671      Length Test Dataset:1504
Accuracy:0.86%
```

Regression Model – Accuracy with split dataset 70-30 is 88%

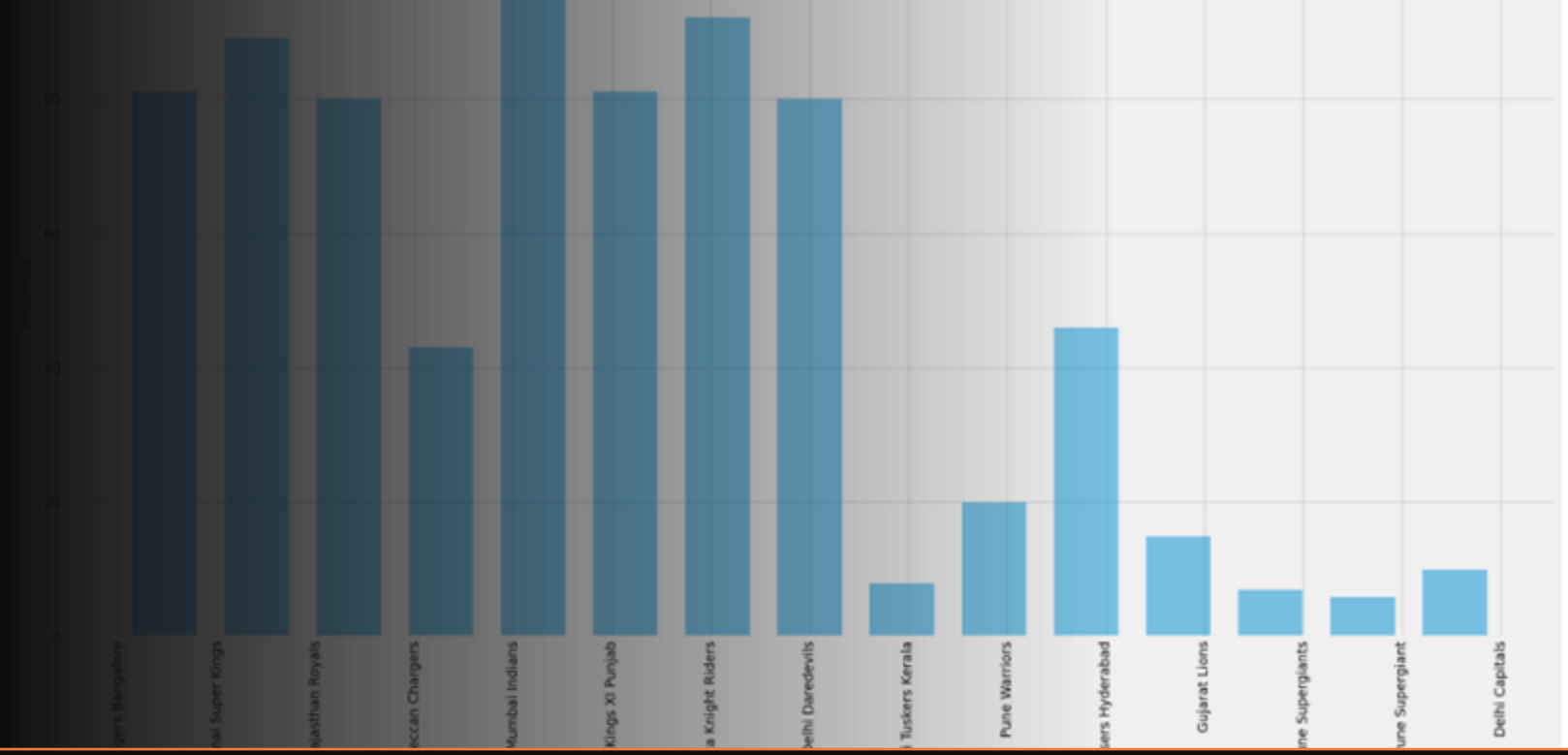
```
Optimization terminated successfully.
         Current function value: 0.473392
         Iterations 8
Intercept              0.143369
teamTotalRuns          0.113018
OpenersTotalRuns       0.014804
oppisitionTotalRuns   -0.119837
dtype: float64
True Positives:196       True Negatives:201      Length Test Dataset:451
Accuracy:0.88%
```

Players with most centuries

Chris Gayle has most number centuries(6)

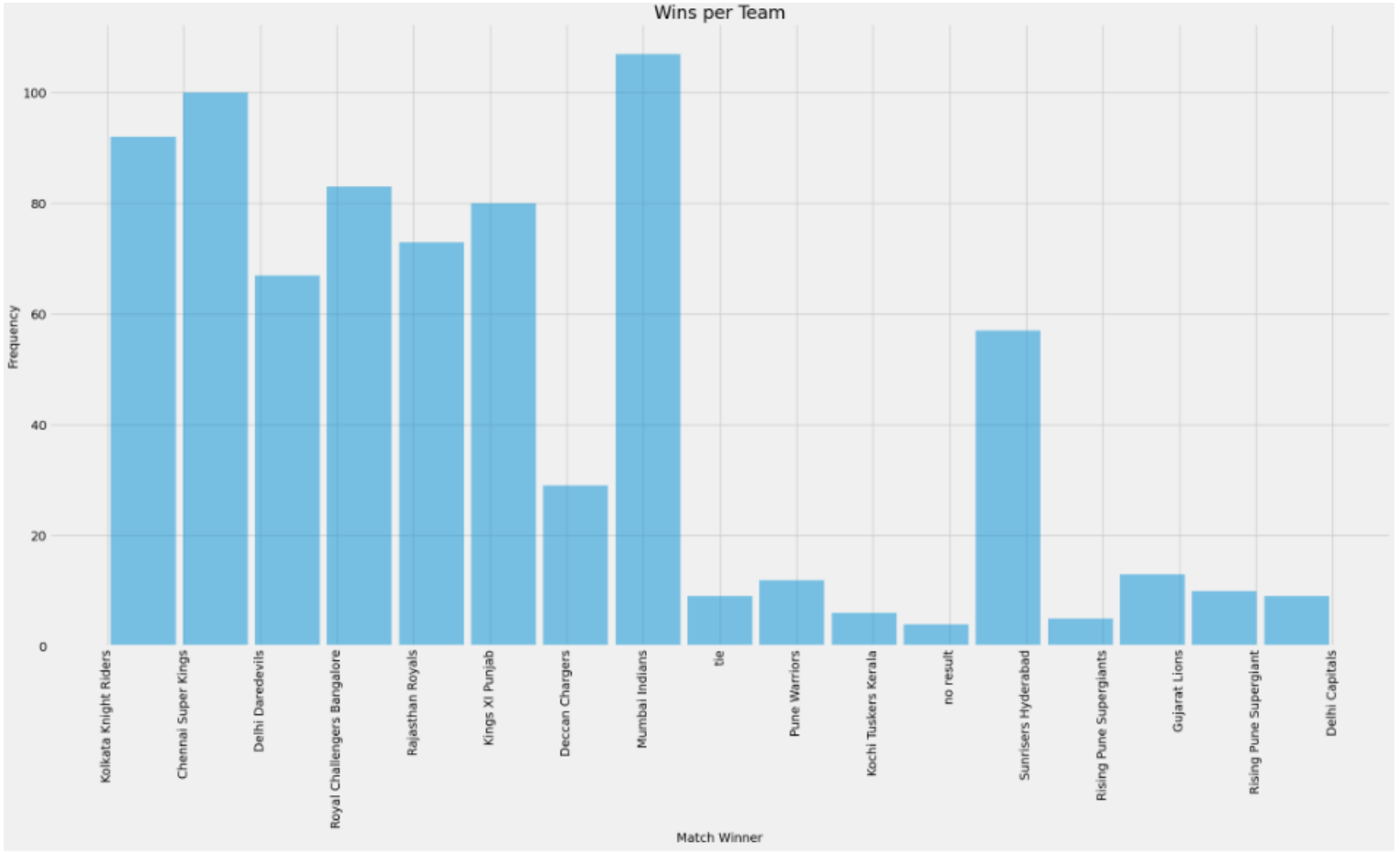


Toss results

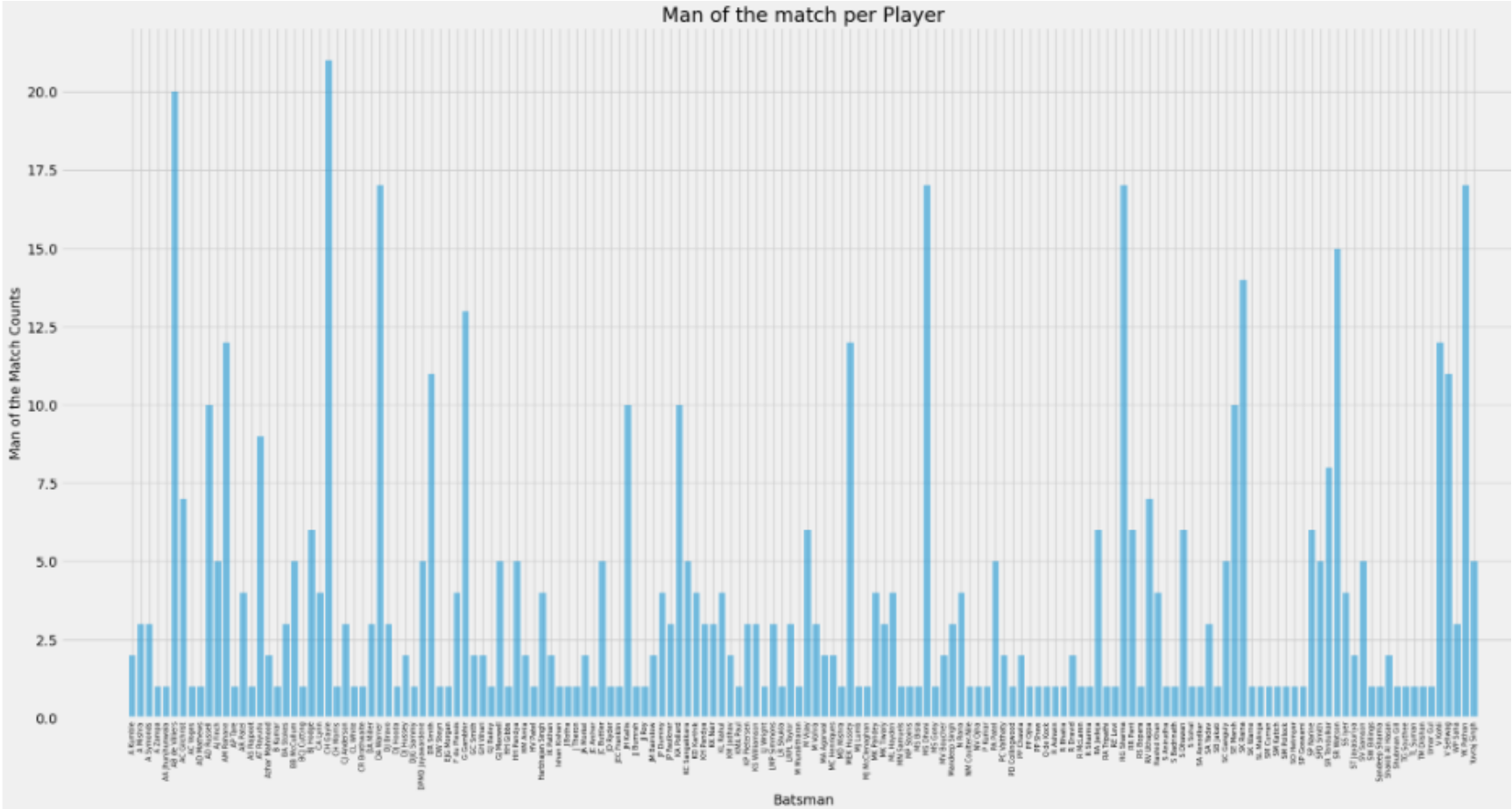


- Mumbai Indians is ahead of all other teams.

Winners across the seasons - Mumbai Indian.



Man of the match - Chris Gayle



What do you feel was missed during the analysis?

- Dataset picked was not current.
- Analysis is happening on past results.
- Individual player performance was limited to IPL tournaments, it could be great if we have considered international and domestic form of the players.

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

• Dataset is batsmen centric, bowlers stats could help better.

• Details about unknown factors like Weather conditions must have been helped.

• Home team variable could have helped to evaluate home advantage.

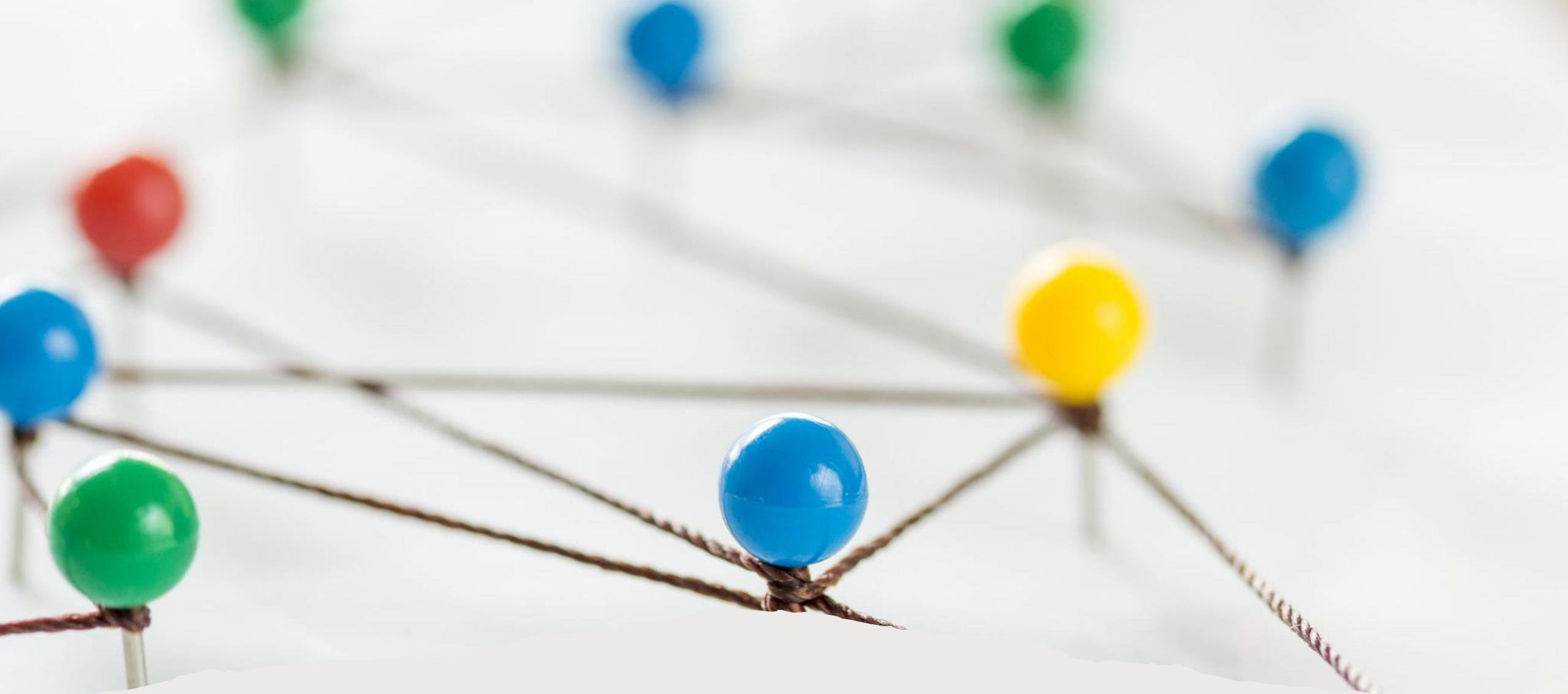
Were there any assumptions made you felt were incorrect?

- Assumptions and hypothesis made initially, was not necessarily supported using data and claims were refuted based on hypothesis testing.

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

Data integration and combining them into to single frame was a big challenge.

Identifying and defining new variables will be the key.



Link to Repository

https://github.com/smunjewar/DSC-530-Data_Exploration_And_Analysis.git

Thank you !