Big Data Analytics on Indian Premier League

Akhil Reddy Kovvuri

G01123172

AIT-580-DL1 ANALYTICS: BIG DATA TO INFORMATION

Dr. Harry J. Foxwell

Spring 2018

Introduction to cricket:

Cricket is a bat and ball game played between the two teams in three Different Formats like T-20, ODI, Test matches. T-20 is 20 over Game which can be played in 3 hours, ODI is a 50 over Game which is played in a single day, whereas Test Matches are played in 5 days with more than 200 overs (Dey , Ghosh and Mondal 2011). There are two innings played in Cricket first innings and second innings, both innings will play the same overs per side, where in first innings one team is batting the other team will bowl to restrict the Batsmen not to score the runs and similarly in the second innings vice versa (Manas 2017)

Who:

Indian Premier League is the T-20 cricket played in India since 10 years. It is the richest sport in India (Bhatia 2017). It is being played with 8 teams, each having 11 players aside. Franchises of Different Teams will auction the players with limited amount of money ($40million USD). Out of 11 players only 4 players from abroad are used to play in a single team. IPL is organized by BCCI (Board of control for cricket in India) and each year it makes $600 million, IPL is like a festival in India every year and it is unmatched following in India (Bhatia 2017).

IPL stores each and every match ball by ball data for the purpose of analyzing the player performances. Every match data is being recorded, this recorded data is analyzed for selecting good performances like batting and bowling. In this project I have collected the data from the Kaggle which provides many data sets for analyzing different real-world problems (Manas 2017).

Data Description:

The collected data has IPL Ball to Ball Data of each match from 2008 to 2017 (Manas 2017). The data consists of 21 attributes and 150461 parameters. The data types of a data are defined as match_id int, inning int, batting_team varchar, bowling_team varchar, over int, ball int, batsman varchar, non_striker varchar, bowler varchar, is_super_over int, wide_runs int, bye_runs int, legbye_runs int, noball_runs int, penalty_runs int, batsman_runs int, extra_runs int, total_runs int, player_dismissed varchar, dismissal_kind varchar, fielder varchar.

As the Data is complex and it can be defined as variety of Data types:

1. “match_id” is Nominal Data type
2. “inning” is Nominal Data type

3. “batting_team” is Nominal Data type
4. “bowling_team” is Nominal Data type
5. “over” is Nominal, ordinal Datatype
6. “ball” is Nominal, ordinal Data type
7. “batsman” is Nominal Data type
8. “non_striker” is Nominal Data type
9. “bowler” is Nominal Data type
10. “is_super_over” is Nominal Data type
11. “wide_runs” Nominal Data type
12. “bye_runs” is Nominal Data type
13. “legbye_runs” is Nominal Data type
14. “noball_runs” is Nominal Data type
15. “penalty_runs” is Nominal Data type
16. “batsman_runs” is Nominal Data type
17. “extra_runs” is Nominal Data type
18. “total_runs” is Nominal, interval, ordinal, ratio data type
19. “player_dismissed” is Nominal Datatype
20. “dismissal_kind” is Nominal Data type
21. “Fielder” is Nominal Data type

Need:

All Indian Premier League Cricket matches between 2008 and 2016 is recorded to evaluate player performances, also the ball by ball data of all the IPL cricket matches till season 9 is recorded (Manas 2017). All the Data recorded includes individual player performances of both batting and bowling. This Data is used for analysis and predict the possible outcomes of the upcoming matches.

There are no privacy related issues with this data because the data is open and any one use for the purpose of analyzing the player performances and also to look what would be outcome of the matches in future.

Questions to be Answered by studying this Data:

- Which IPL Team scored Highest Number of Runs in Batting
- Which IPL Team Taken Highest Number of Wickets in Bowling

- Which IPL Team has scored highest number of Boundaries
- Top Batsmen's in IPL
- Top Bowler's in IPL

Requirements and Tools used:

In this project I have used SQL, R, Tableau, Python to study the data. SQL is used for Designing Schema, R is Used for Descriptive Statistics, Tableau is used for Visualizations and python to analyze the Best Batsman and Bowling performances in IPL.

Metadata Using SQL:

The IPL (Ball by Ball deliveries) data is inserted into SQL database (workbench) to identify the attribute types (String and Integer). Below screenshots explains the metadata and data description using SQL queries. Figure 1 shows the insert commands into the SQL server, Figure 2 shows the data type of each attribute and Figure 3 shows the ball by ball that inserted in to the table.

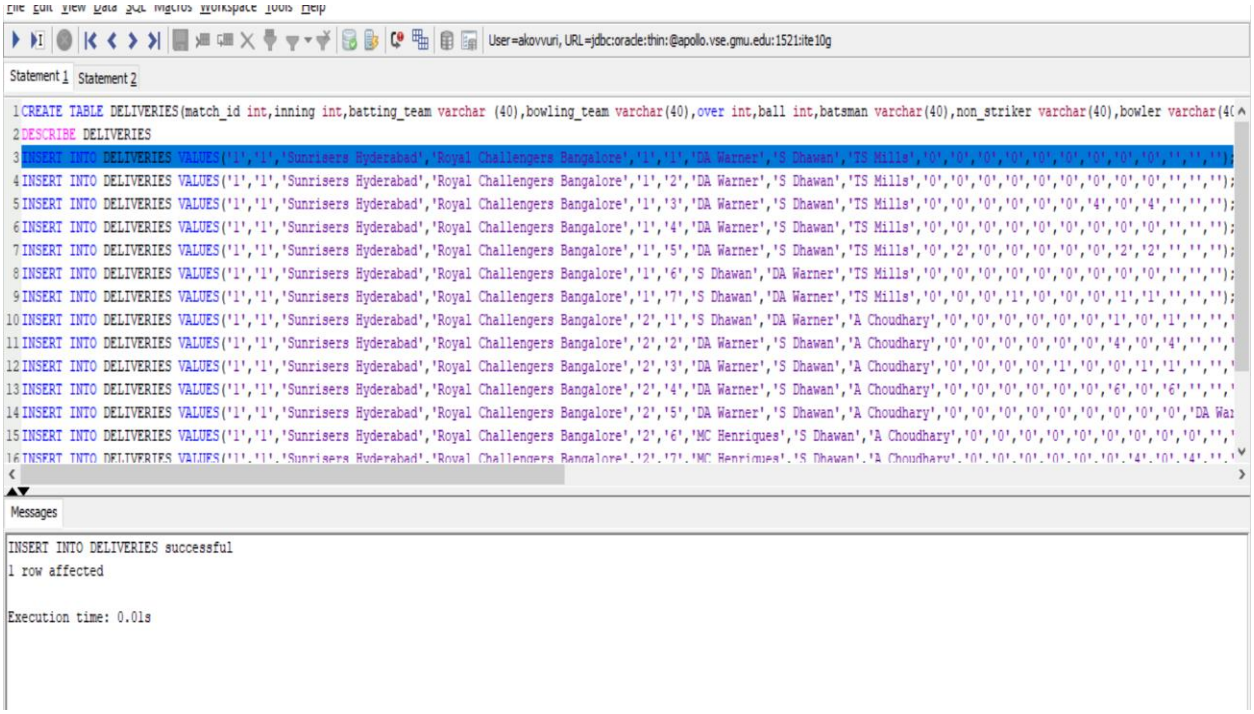

Fig 1: Inserted values in the Table

AKOVVURI.DELIVERIES (TABLE) Messages

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DEFAULT	AUTOINCREMENT	COMPUTED	REMARKS	POSITION
BOWLER	VARCHAR2(40 Byte)	NO	YES		NO	NO		9
IS_SUPER_OVER	NUMBER	NO	YES		NO	NO		10
WIDE_RUNS	NUMBER	NO	YES		NO	NO		11
BYE_RUNS	NUMBER	NO	YES		NO	NO		12
LEGBYE_RUNS	NUMBER	NO	YES		NO	NO		13
NOBALL_RUNS	NUMBER	NO	YES		NO	NO		14
PENALTY_RUNS	NUMBER	NO	YES		NO	NO		15
BATSMAN_RUNS	NUMBER	NO	YES		NO	NO		16
EXTRA_RUNS	NUMBER	NO	YES		NO	NO		17
TOTAL_RUNS	NUMBER	NO	YES		NO	NO		18
PLAYER_DISMISSED	VARCHAR2(20 Byte)	NO	YES		NO	NO		19
DISMISSAL_KIND	VARCHAR2(40 Byte)	NO	YES		NO	NO		20
FIELDER	VARCHAR2(40 Byte)	NO	YES		NO	NO		21

Fig 2: Data types of Attributes

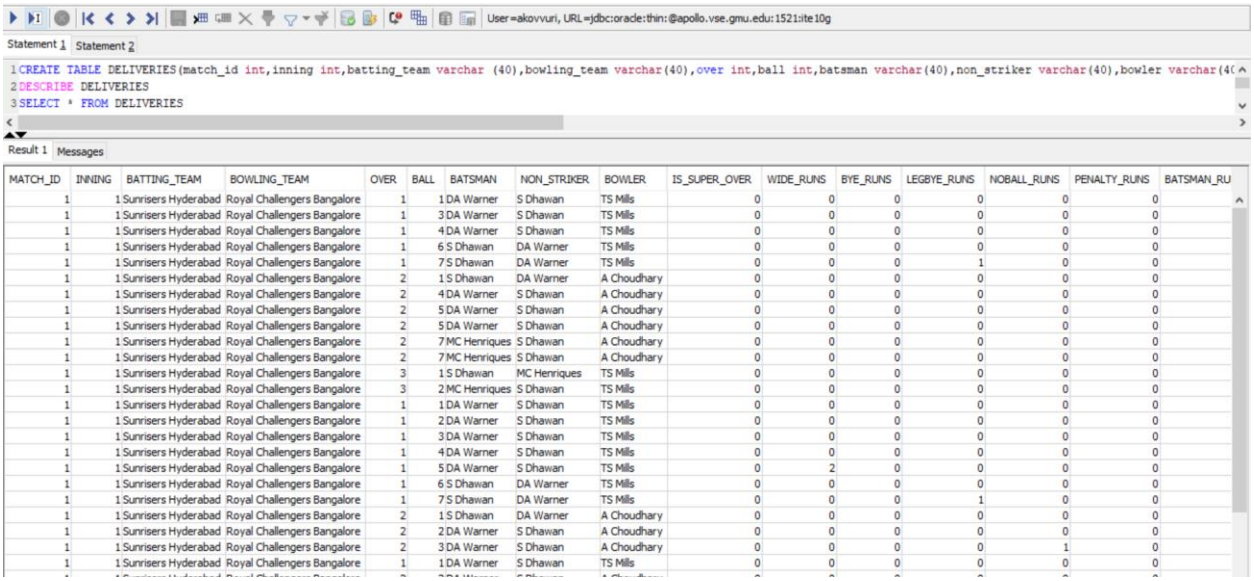

Fig 3: Displaying the Table Records

Exploring the Dataset with Python:

I have explored the dataset by python, which finds the Best Top Batsman and Best Top Bowlers. Figure 4 shows the Top 5 Batsman in IPL from 2008 to 2017 whereas figure 5 shows the Top 10 Bowlers in IPL.

```
In [1]: import pandas as pd

In [2]: import numpy as np

In [3]: Bat= pd.read_csv("BatsmanCareerRuns.csv")

In [4]: Bat.head()
```

Out[4]:

	Batsman	CareerRuns
0	A Ashish Reddy	196
1	A Chandila	7
2	A Chopra	75
3	A Choudhary	20
4	A Flintoff	57

```
In [13]: print(Bat[['CareerRuns','Batsman']].loc[Bat['CareerRuns']>3000])

     CareerRuns    Batsman
137        3433  G Gambhir
324        3274  RG Sharma
347        3005   S Dhawan
383        3369   SK Raina
431        3494    V Kohli
```

Fig 4: Best Batsman of IPL

```
In [1]: import pandas as pd

In [2]: import numpy as np

In [3]: Bowl=pd.read_csv("BowlersTotalWickets.csv")

In [4]: Bowl.head()

Out[4]:
```

	Bowler	TotalWickets
0	A Ashish Reddy	19.0
1	A Chandila	11.0
2	A Choudhary	5.0
3	A Flintoff	2.0
4	A Kumble	49.0

```
In [6]: Bowl.shape

Out[6]: (356, 2)

In [7]: print(Bowl[['TotalWickets','Bowler']].loc[Bowl['TotalWickets']>100])

     TotalWickets           Bowler
5           142.0         A Mishra
7           121.0          A Nehra
50          117.0          B Kumar
88          137.0         DJ Bravo
115         136.0  Harbhajan Singh
218         102.0          P Kumar
231         133.0        PP Chawla
234         110.0          R Ashwin
244         125.0    R Vinay Kumar
302         170.0       SL Malinga
309         109.0        SP Narine
334         107.0         UT Yadav
355         119.0          Z Khan
```

Fig 5: Best Top Bowlers in IPL

Visualizations using R and Tableau:

Below are the visualizations made through R and Tableau by plotting the different attributes. From Figure 6 I have analyzed the team with highest number of boundaries scored till now. From Figure 7, I have analyzed the 14 teams Best Batting Performances in the IPL from 2008 to 2017 whereas from figure 8, I have analyzed the Best Bowling Team Performances in the IPL

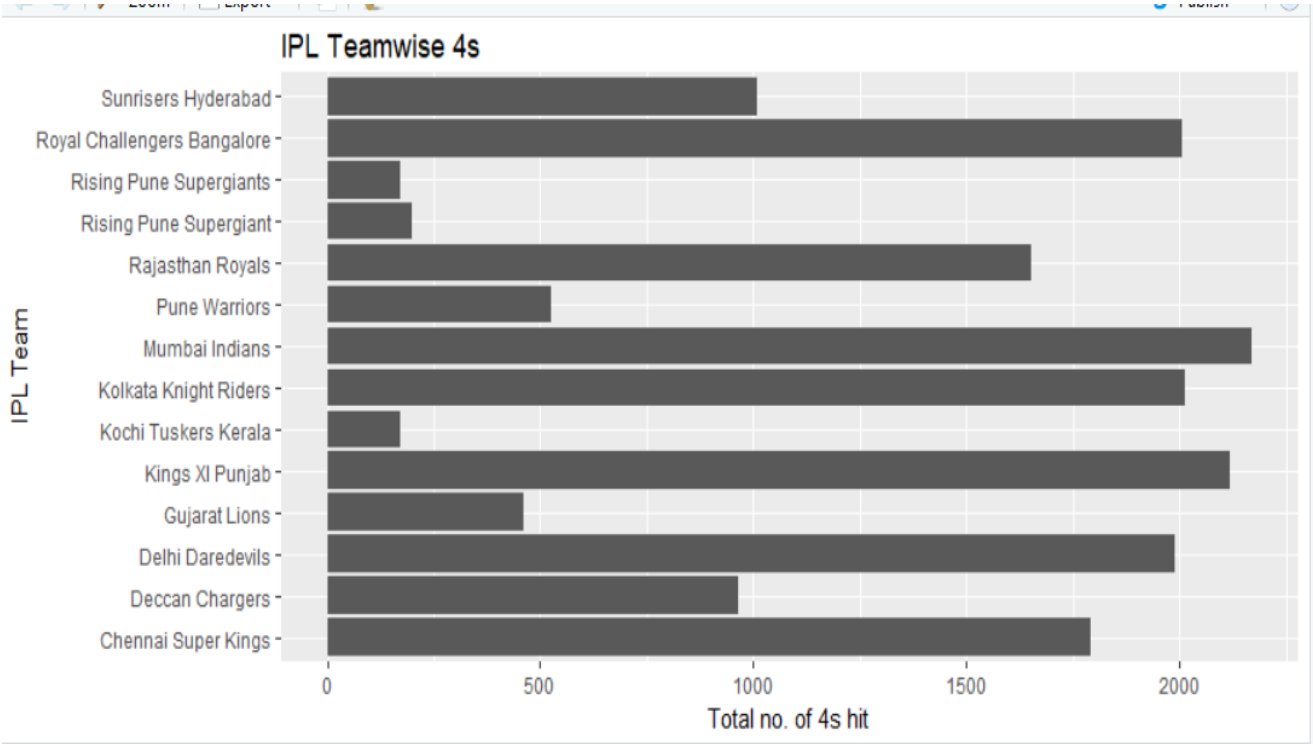

Fig 6: Team with Highest Boundaries

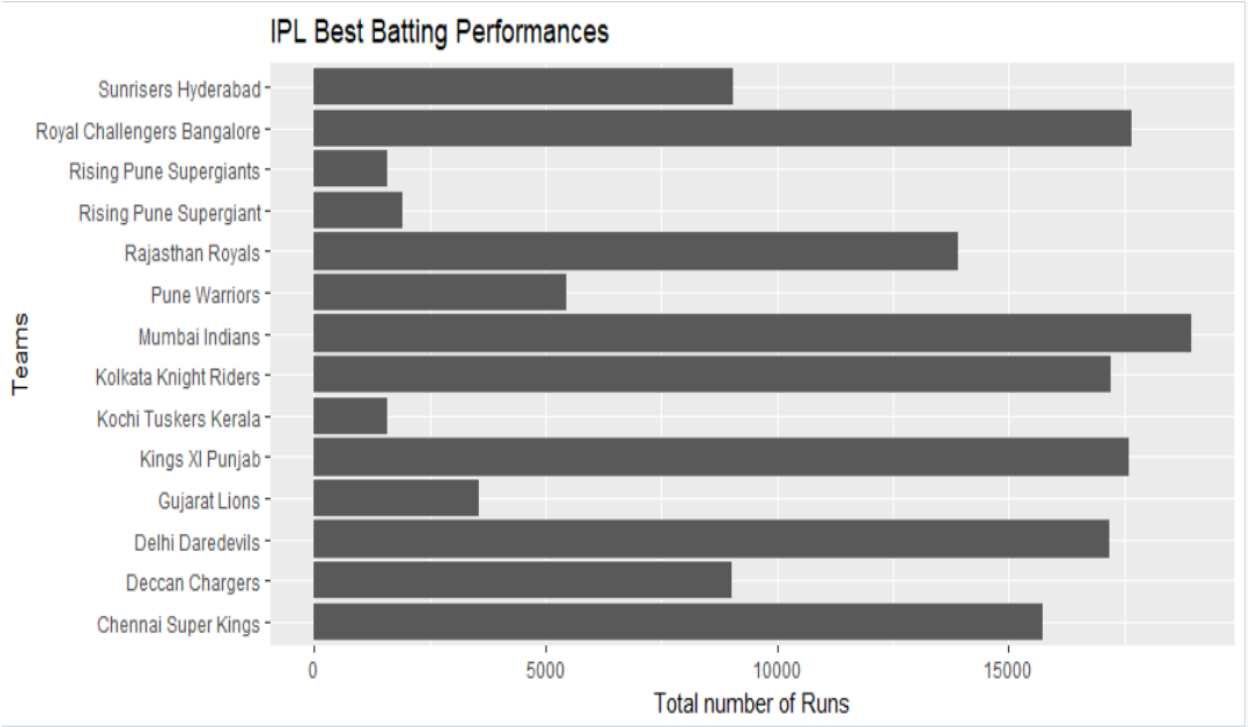

Fig 7: Team with Best Batting Performances in IPL

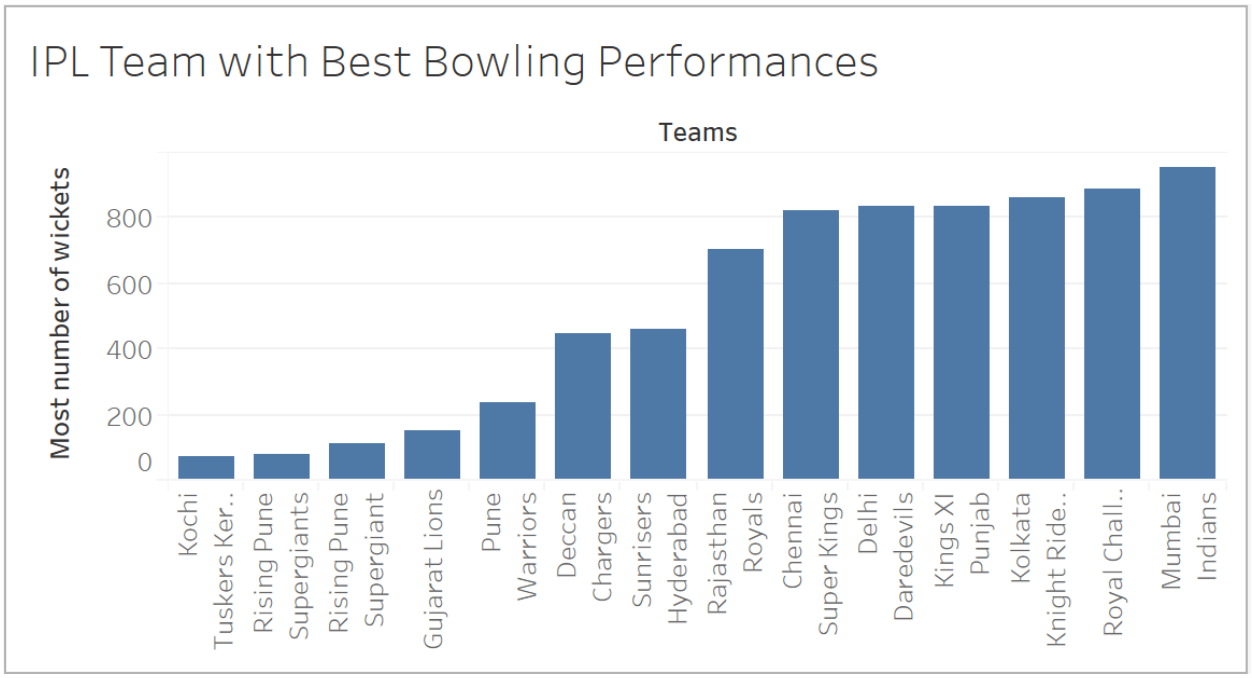

Fig 8: Team with Best Bowling Performances in IPL

Arithmetic Calculations:

Below are some arithmetic calculations that have been performed on the data set using R. From the figure 9 and 10, I can see the Average, Median and maximum of the runs scored and the best batsman and Bowler attributes of players played in the IPL.

```
> IPL<-read.csv("C:/Users/akhil/Downloads/ipl/deliveries.csv")
> summary(IPL)
    match_id         inning                          batting_team
 Min.   :  1.0   Min.   :1.000   Mumbai Indians              :18943
 1st Qu.:161.0   1st Qu.:1.000   Royal Challengers Bangalore:17678
 Median :319.0   Median :1.000   Kings XI Punjab             :17594
 Mean   :318.3   Mean   :1.482   Kolkata Knight Riders       :17229
 3rd Qu.:476.0   3rd Qu.:2.000   Delhi Daredevils            :17185
 Max.   :636.0   Max.   :4.000   Chennai Super Kings         :15754
                                 (Other)                     :46077
                     bowling_team        over             ball
 Mumbai Indians              :18879   Min.   : 1.00   Min.   :1.000
 Royal Challengers Bangalore:17920   1st Qu.: 5.00   1st Qu.:2.000
 Kolkata Knight Riders       :17411   Median :10.00   Median :4.000
 Kings XI Punjab             :17392   Mean   :10.14   Mean   :3.616
 Delhi Daredevils            :17099   3rd Qu.:15.00   3rd Qu.:5.000
 Chennai Super Kings         :15562   Max.   :20.00   Max.   :9.000
 (Other)                     :46197
       batsman           non_striker                bowler
 V Kohli   :  3494   G Gambhir :  3635   Harbhajan Singh:  2989
 G Gambhir :  3433   SK Raina  :  3483   A Mishra       :  2703
 SK Raina  :  3369   V Kohli   :  3351   SL Malinga     :  2694
 RG Sharma :  3274   RG Sharma :  3306   P Kumar        :  2637
 S Dhawan  :  3005   S Dhawan  :  3248   PP Chawla      :  2594
 RV Uthappa:  2960   RV Uthappa:  2848   R Ashwin       :  2359
 (Other)   :130925   (Other)   :130589   (Other)        :134484
 is_super_over        wide_runs         bye_runs         legbye_runs
 Min.   :0.0000000   Min.   :0.0000   Min.   :0.000000   Min.   :0.00000
 1st Qu.:0.0000000   1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000
 Median :0.0000000   Median :0.0000   Median :0.000000   Median :0.00000
 Mean   :0.0005383   Mean   :0.0375   Mean   :0.004885   Mean   :0.02223
 3rd Qu.:0.0000000   3rd Qu.:0.0000   3rd Qu.:0.000000   3rd Qu.:0.00000
 Max.   :1.0000000   Max.   :5.0000   Max.   :4.000000   Max.   :5.00000
```

Fig 9: Descriptive statistics of the data

```
 noball_runs        penalty_runs       batsman_runs      extra_runs
Min.   :0.00000   Min.   :0.0e+00   Min.   :0.000   Min.   :0.00000
1st Qu.:0.00000   1st Qu.:0.0e+00   1st Qu.:0.000   1st Qu.:0.00000
Median :0.00000   Median :0.0e+00   Median :1.000   Median :0.00000
Mean   :0.00434   Mean   :6.6e-05   Mean   :1.222   Mean   :0.06902
3rd Qu.:0.00000   3rd Qu.:0.0e+00   3rd Qu.:1.000   3rd Qu.:0.00000
Max.   :5.00000   Max.   :5.0e+00   Max.   :6.000   Max.   :7.00000

  total_runs      player_dismissed  dismissal_kind
Min.   :0.000             :143022         :143022
1st Qu.:0.000   SK Raina  :   134   caught :  4373
Median :1.000   G Gambhir :   131   bowled :  1382
Mean   :1.291   RG Sharma :   129   run out:   755
3rd Qu.:1.000   RV Uthappa:   128   lbw    :   455
Max.   :7.000   V Kohli   :   118   stumped:   243
                (Other)   :  6798   (Other):   230
          fielder
              :145091
KD Karthik    :   127
MS Dhoni      :   126
RV Uthappa    :   115
AB de Villiers:   101
SK Raina      :    96
(Other)       :  4804
```

Fig: Descriptive Statistics of the Dataset.

Conclusions:

From the above analysis, I have analyzed the Best Batsmen's, Best Bowlers, Best Batting Team, Best Bowling Team. From the analysis it is clear that Mumbai Indians is the best Batting and Bowling Team. Also, The Top Best Batsmen's are Gambhir, Sharma, Dhawan, Virat kohli and suresh Raina, whereas the Top Best Bowlers are A. Mishra, A. Nehra, B. Kumar, D.J. Bravo, Harbajan singh, P. Kumar, PP. Chawla, R. Ashwin, R. Vinay Kumar, Malinga, Yadav, Narine, Zaheer Khan.

Define Terms:

Non-striker: Non-striker is the batsmen who stands at the other end to run between the wickets, when the other batsmen is playing.

Is_Super_over: Super over is played when the match is draw or tie (scores are equal).

Leg_bye: Leg bye is the extra runs scored without hitting the ball with bat.

Dismissal_kind: Dismissal is like how the player is Out (Bowled or Caught or stumped or runout).

References

Bhatia, Gauri. 2017. *The Richest sport in India just Getting richest in Richer.* Indian-premier-league-cricket, https://www.cnbc.com.

Dey , Pabitra Kumar, Dipendra Nath Ghosh, and Abhoy Chand Mondal. 2011. "A MCDM approach for Evaluating Bowlers Performances in IPL." *Journal of Emerging Trends in Computing and Information Sciences.*

Manas. 2017. *Indian Premier League.* https://www.kaggle.com/manasgarg/ipl.