**IPL ANALYSIS**

Probability of Winning When Batting First

**Sougata Khan**

**B.Sc(Hons)-Statistics**

**Semester–5**

**Enrollment Number : A90555919014**

***Introduction:*** For any bivariate data the calculation of statistics i.e. mean, median, mode, quartiles etc is very complicative. That’s why we use Correlation and Regression. Correlation says if there is any relation between two data set and regression finds the relation between them.

**What is Regression**: Regression to mean is all about how data evens out. It basically states that if a variable is extreme the first time you measure it, it will be closer to the average the next time you measure it. In technical terms, it describes how a random variable that is outside the norm eventually tends to return to the norm.

**Definition:** The word ‘Regression’ is used to demote the estimation or prediction of the average value of one variable for the specified value of other variables.

In regression analysis the independent variable is called Regressor or Predictor or Explainator. While the dependent variable is called Regressed, or Explained.



This is the equation of simple linear regression and



This is the equation of multiple linear regression.

We can’t use the same equations to predict the probability because-

1. The probability of the variable lies between 0 & 1
2. Probability doesn’t vary linearly

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable.

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression(or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a *logit*, from *logistic unit*, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

In a binary logistic regression model, the dependent variable has two levels (categorical). Outputs with more than two values are modeled by multinomial logistic regression and, if the multiple categories are ordered, by ordinal logistic regression (for example the proportional odds ordinal logistic model). The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier. The coefficients are generally not computed by a closed-form expression, unlike linear least square.

The logistic regression as a general statistical model was originally developed and popularized primarily by Joseph Berkson beginning in Berkson (1944), where he coined "logit"

*Logistic regression analysis* studies the association between a categorical dependent variable and a set ofindependent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

**The Link function**

 { where p=probability of success}

Or, 

[**Binary logistic regression**](https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html#id18)

Say we’re given data on student exam results and our goal is to predict whether a student will pass or fail based on number of hours slept and hours spent studying. We have two features (hours slept, hours studied) and two classes: passed (1) and failed (0).Graphically we could represent our data with a scatter plot.

1. [**Sigmoid activation**](https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html#id19)

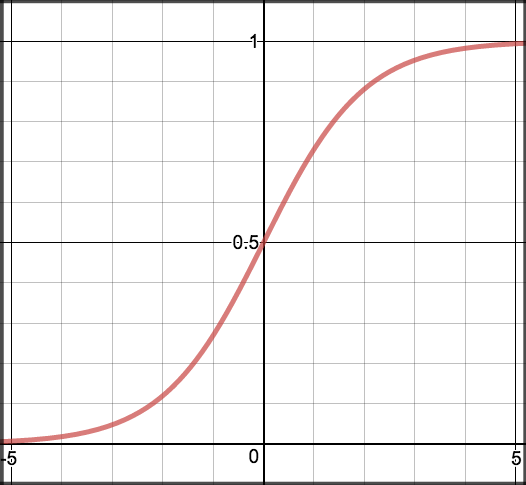
In order to map predicted values to probabilities, we use the [sigmoid](https://ml-cheatsheet.readthedocs.io/en/latest/activation_functions.html#activation-sigmoid) function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.

Math

S(z)=11+e−zS(z)=11+e−z

Note

* s(z)s(z) = output between 0 and 1 (probability estimate)
* z = input to the function (your algorithm’s prediction e.g. mx + b)
* e = base of natural log

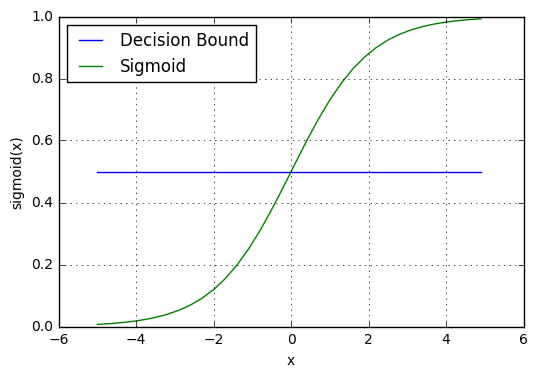


# [Decision boundary](https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html#id20)

Our current prediction function returns a probability score between 0 and 1. In order to map this to a discrete class (true/false, cat/dog), we select a threshold value or tipping point above which we will classify values into class 1 and below which we classify values into class 2.

p≥0.5, class=1 p<0.5, class=0 p≥0.5, class=1 p<0.5, class=0

For example, if our threshold was .5 and our prediction function returned .7, we would classify this observation as positive. If our prediction was .2 we would classify the observation as negative. For logistic regression with multiple classes we could select the class with the highest predicted probability.



Model:

Let us try to understand logistic regression by considering a logistic model with given parameters, then seeing how the coefficients can be estimated from data. Consider a model with two predictors, x1 and x2 one binary (Bernoulli) response variable Y , which we denote p=P(Y-1) . We assume a linear relationship between the predictor variables and the log-odds of the event that Y=1. This linear relationship can be written in the following mathematical form (where *ℓ* is the log-odds, b is the base of the logarithm, and  are parameters of the model):



We can recover the [odds](https://en.wikipedia.org/wiki/Odds) by exponentiating the log-odds:

.

By simple algebraic manipulation, the probability that Y=1 is



The above formula shows that once are fixed, we can easily compute either the log-odds that Y=1 for a given observation, or the probability that Y=0 for a given observation. The main use-case of a logistic model is to be given an observation  , and estimate the probability p that Y=1 . In most applications, the base b of the logarithm is usually taken to be [*e*](https://en.wikipedia.org/wiki/E_(mathematical_constant)). However in some cases it can be easier to communicate results by working in base 2, or base 10.

We consider an example with b=10, and coefficients , . To be concrete, the model is



where p is the probability of the event that Y=1 .

This can be interpreted as follows:

*  is the y-intercept. It is the log-odds of the event that Y=1 , when the predictors x1=x2=0 . By exponentiating, we can see that when x1=x2=0 the odds of the event that Y=1 are 1-to-1000, Similarly, the probability of the event that Y=1 when x1=x2=0 can be computed as .1/(1000+1)=1/1001
*  means that increasing x1 by 1 increases the log-odds by 1 . So if x1  increases by 1, the odds that Y=1 increase by a factor of 101 . Note that the probability of Y=1 has also increased, but it has not increased by as much as the odds have increased.
* means that increasing x2 by 1 increases the log-odds by 2 . So if x2 increases by 1, the odds that Y=1 increases by a factor of 102 .Note how the effect of x2 on the log-odds is twice as great as the effect of x1 , but the effect on the odds is 10 times greater. But the effect on the probability of Y=1 is not as much as 10 times greater, it's only the effect on the odds that is 10 times greater.

In order to estimate the parameters  from data, one must do logistic regression.

ABSTRACT:

This study is novel in that it used both pre-match variables (home-ground advantage) and real-time measures (e.g., how many runs were scored in the powerplay) in a statistical context to classify match results. The results can be used to adapt in-game tactics to maximize advantages of batsmen in favorable contexts.

INTRODUCTION:

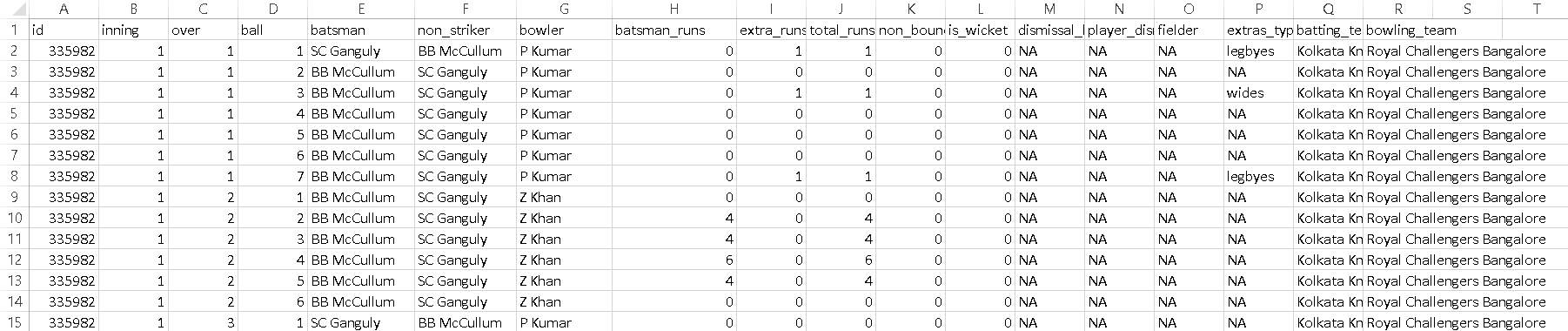
If we want to find the probability of winning of a team in IPL when batting first, then we have to look after the previous matches of that team. Using the past data of every matches we can do that. Here comes “Regression”. The word ‘Regression’ is used to denote the estimation or prediction of the average value of one variable for the specified value of other variables. And as we are going to find probabilities, that‘s why here we shall use “Logistic Regression”. Logistic regression is **a statistical analysis method used to predict a data value based on prior observations of a data set**. ... A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

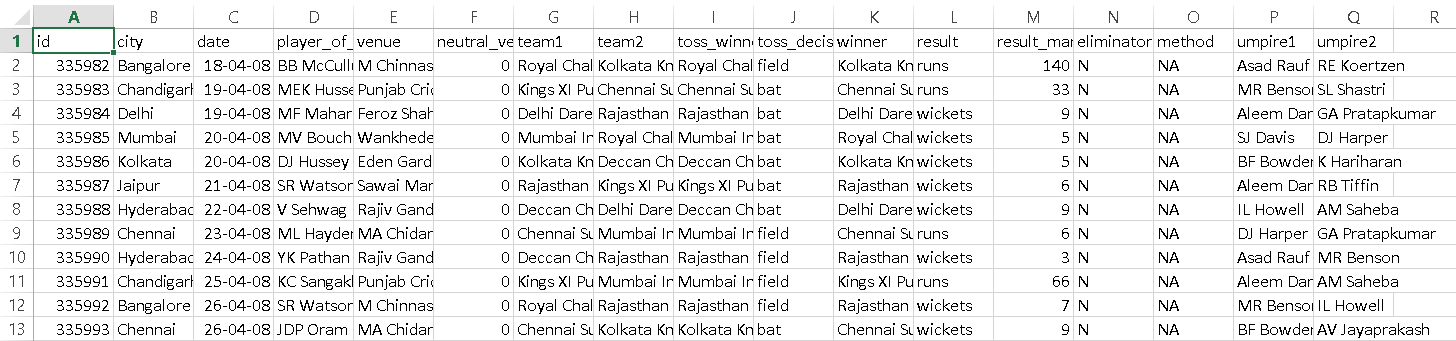
The variable “winning” depends on many other variables in IPL. Like : Total runs, Number of runs scored in the first six overs of the innings, Number of dot balls, one runs, two runs, three runs, boundaries, sixers in the innings, Number of wickets in the innings, Number of wickets in the powerplay, Home ground advantage.

OBJECTIVES:

Every team has dream of winning the final cup or trophy. The dream becomes their goal and that goal is achieved by planned action for unfavourable scenarios. Our purpose is to use data from the Indian Premier League cricket tournament to predict match outcomes based on events occurring in the first inning of a match.

DATA:





Two excel files data have been used to do this project. One is IPL ball by ball data and another one is match data.

**[Data Source :** https://www.kaggle.com/patrickb1912/ipl-complete-dataset-20082020**]**

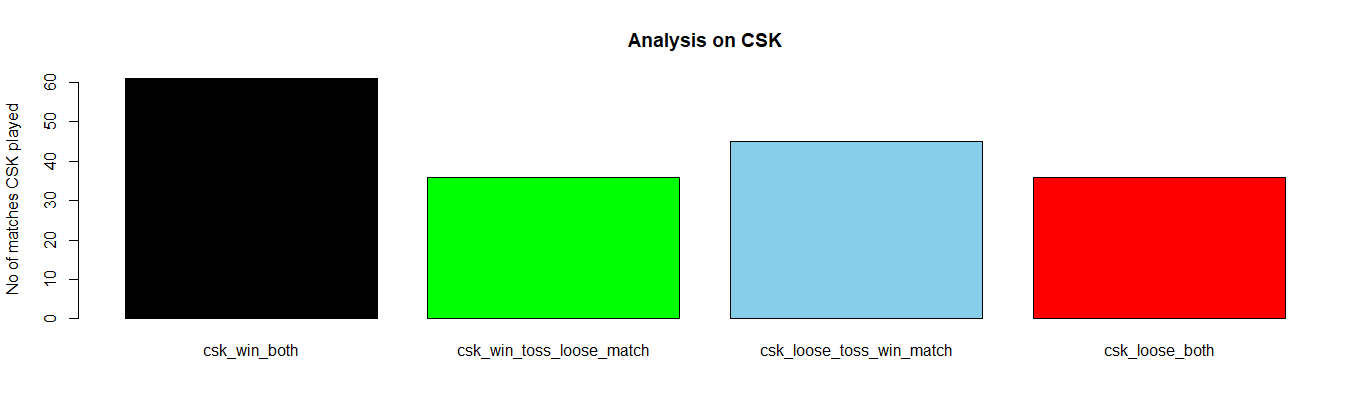
METHODOLOGY:

All regular season and finals matches from the inaugural 2008 season up to the recently completed 2020 season were entered into pre-processing. For all analyses, matches that were ties or abandoned were removed (i.e., only wins and losses were used), and any match that was shortened (i.e., due to inclement weather conditions) was also excluded. This resulted in a final sample size of 780 matches.

At first I have imported two datasets in R-programming. Then I did data manipulation i.e. I analyzed the data and found the answers of some important questions. Next, I built a logistic regression model to find the probability of winning when batting first.

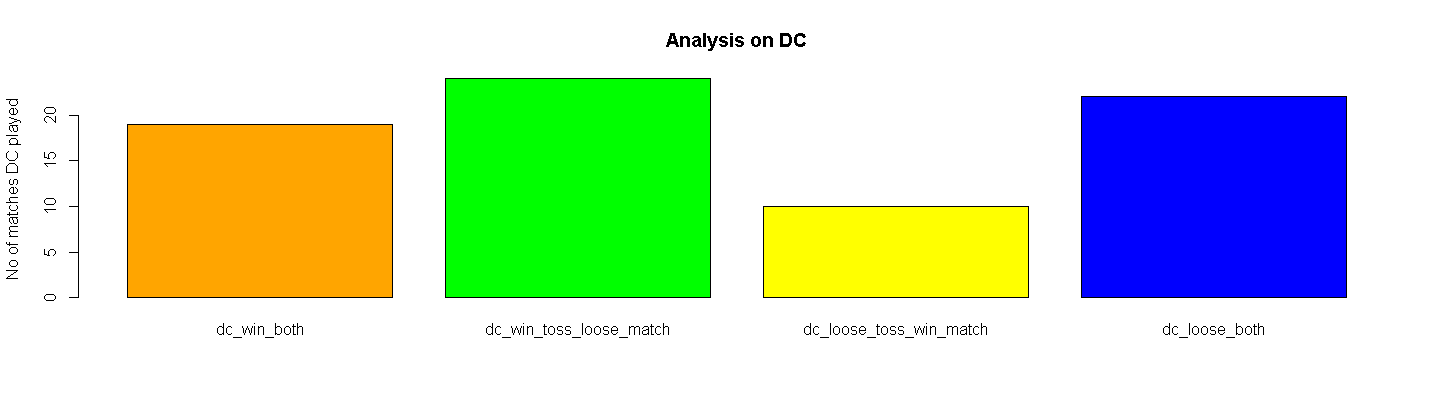
RESULTS:

**Has Toss-winning helped in winning matches?**



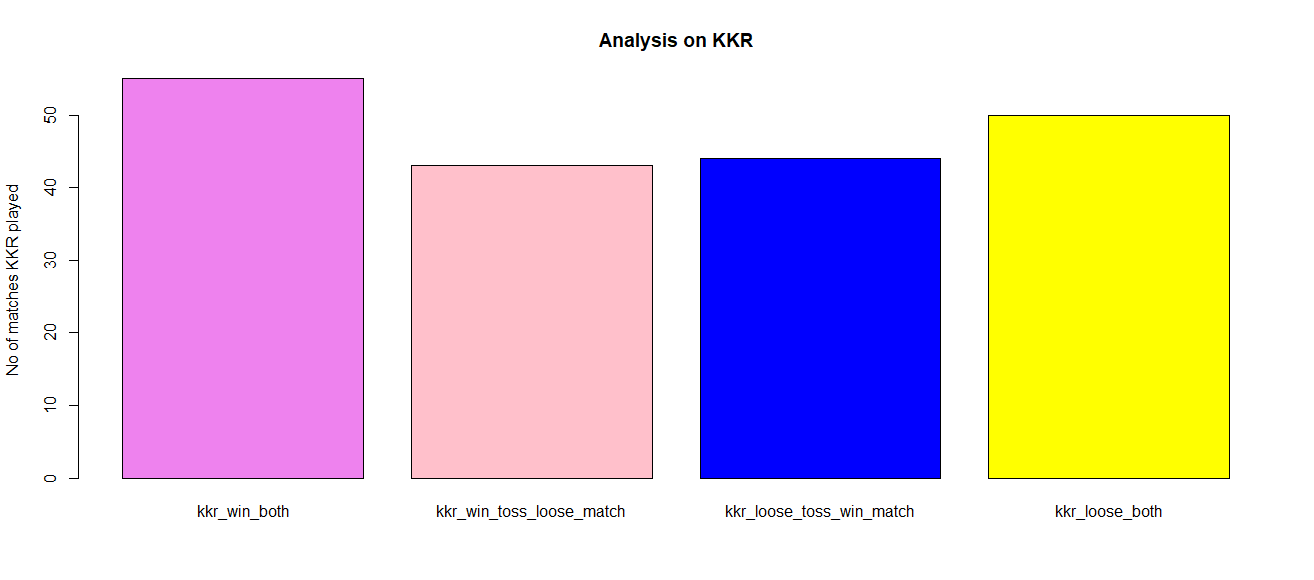
# 61 times CSK wins both in toss and match

* For Chennai Super Kings team, yes. Toss winning helped in winning matches too.



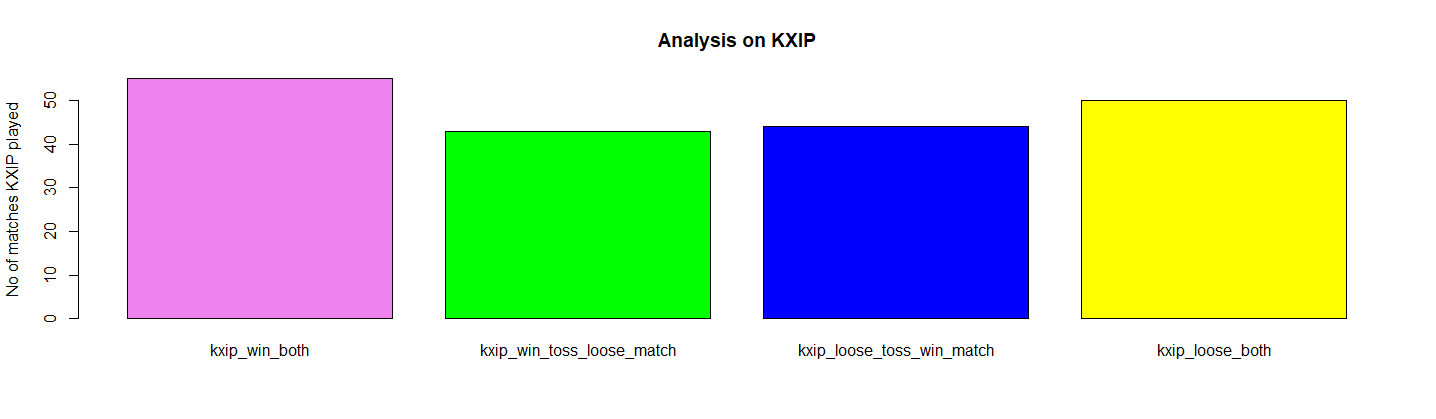
# 19 times DC wins both in toss and match

* For Deccan Chargers team, no. Toss winning didn’t help in winning matches. They won toss in many matches but lost in match



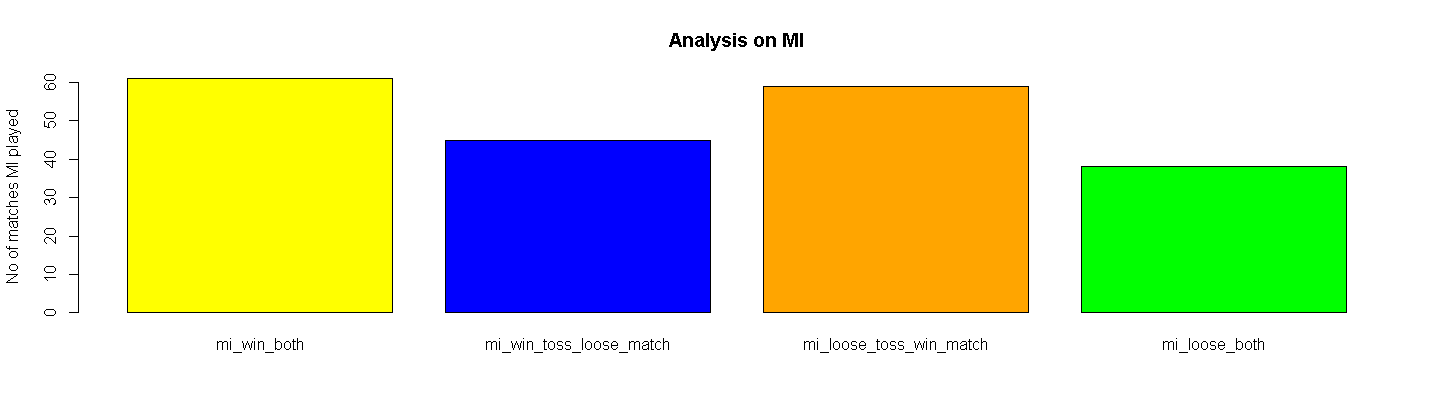
# 55 times KKR wins both in toss and match

* For Kolkata Knight Riders team, yes. Toss winning helped in winning matches too.



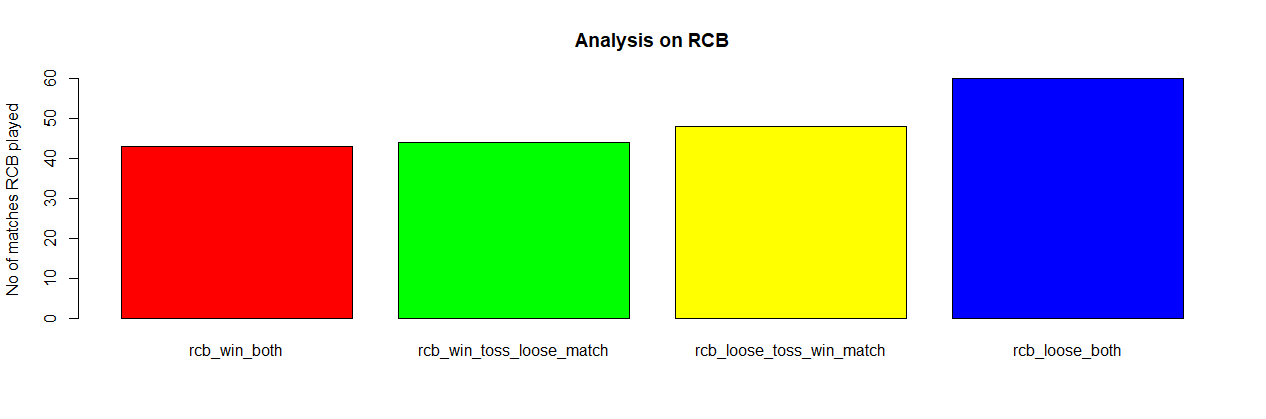
# 36 times KXIP wins both in toss and match

* For Kings XI Punjab team, yes. Toss winning helped in winning matches.



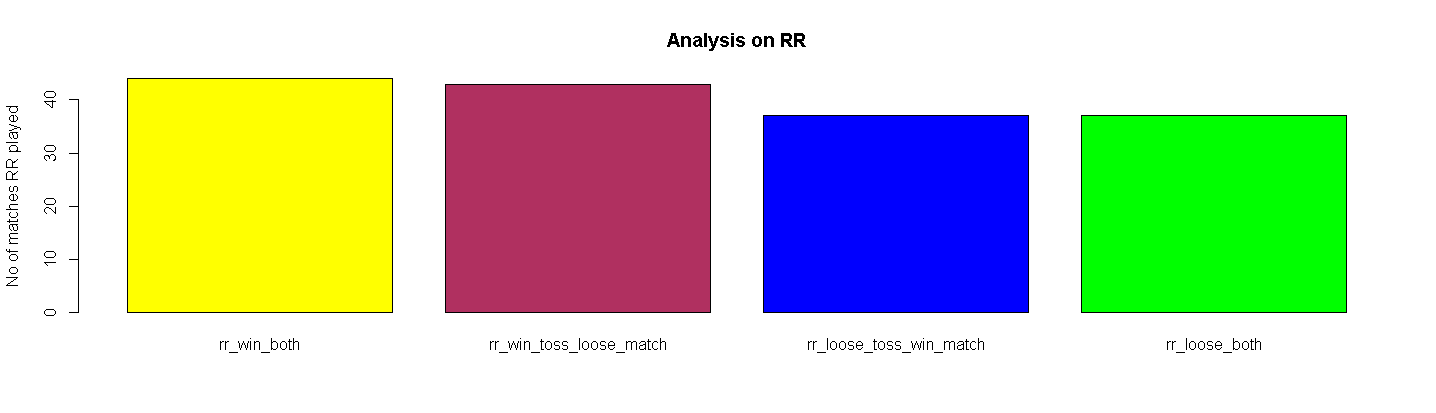
# 61 times MI wins both in toss and match

* But for Mumbai Indians team, it can’t be told. The no of times toss winning helped in winning matches and no of times loosing in toss but won the match are approximately equal.



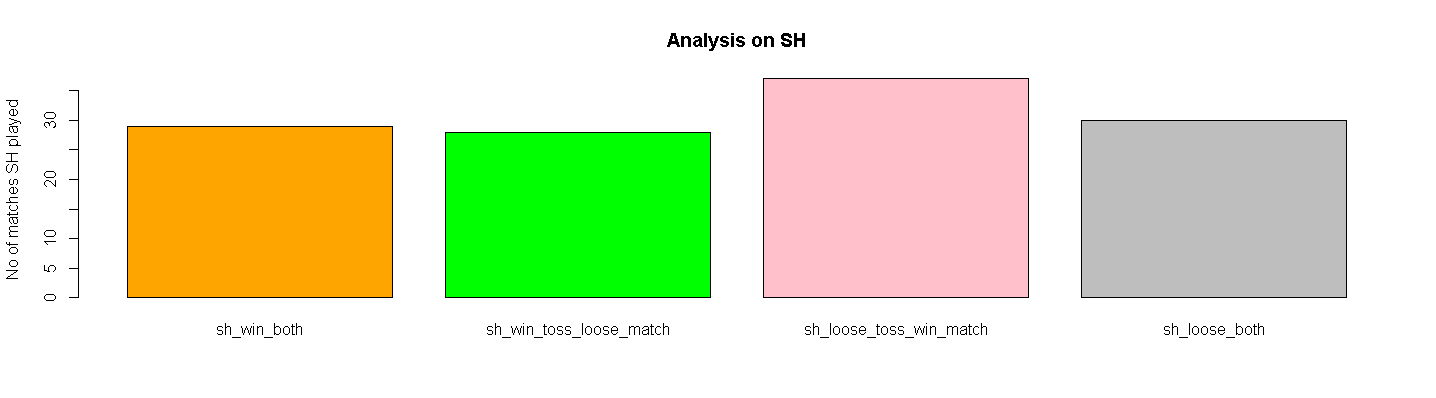
# 43 times RCB wins both in toss and match

* For Royal Challengers Bangalore team, it can’t be told. The no of times toss winning helped in winning matches and no of times winning in toss but lost the match are approximately same, and graph shows that if RCB looses in toss, it may loose match too.



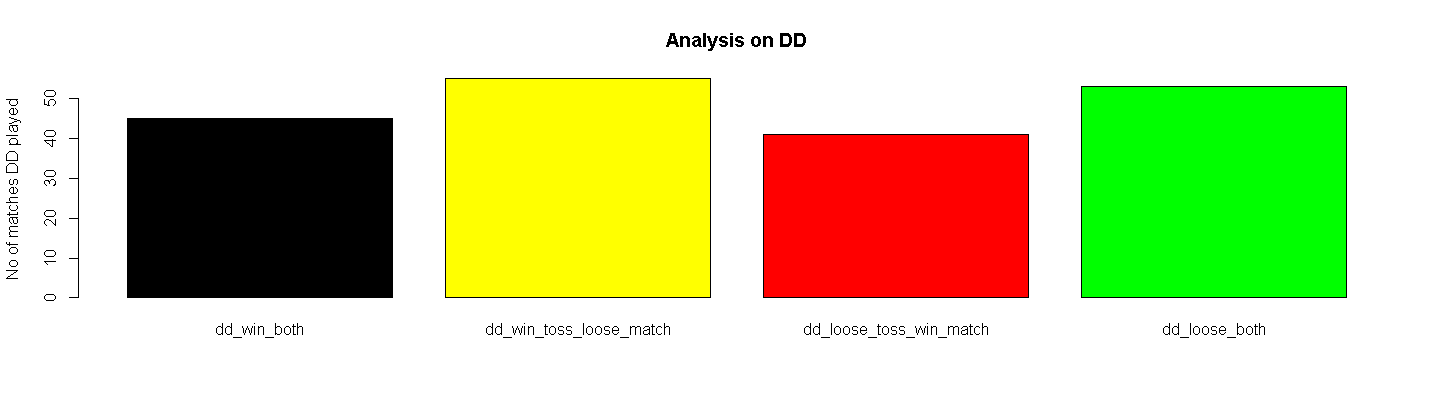
# 44 times RR wins both in toss and match

* But for Rajasthan Royals team, yes. Toss winning helped in winning matches.



# 29 times SH wins both in toss and match

* For Sunrisers Hyderabad team, no. Many Times they lost in toss but won the match.



# 45 times DD wins both in toss and match

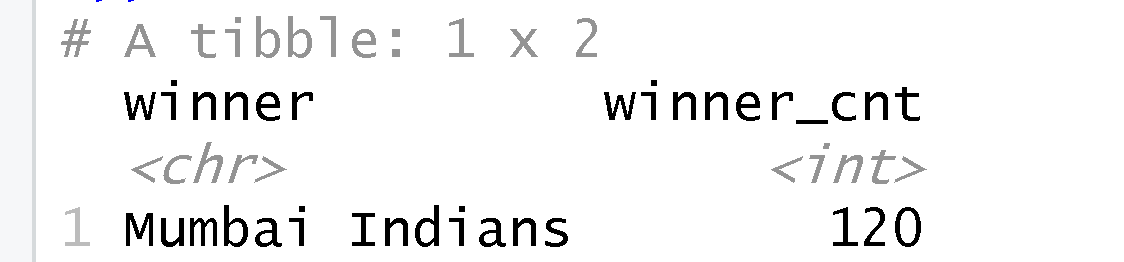
* But for Delhi Daredevils (Delhi Capitals) team, no. Because graph shows, many times they won the toss but lost the match.

**Team with most win or Most successful team**

match%>%

group\_by(winner)%>%

summarise(winner\_cnt=n())%>%filter(winner\_cnt==max(winner\_cnt))



* Team Mumbai Indians is the most successful team in IPL history because it has won highest matches (120 matches).

**Most man of the match winner**

match%>%

filter(result != 'no result') %>% group\_by(player\_of\_match) %>%

summarise(win = n()) %>% arrange(desc(win))

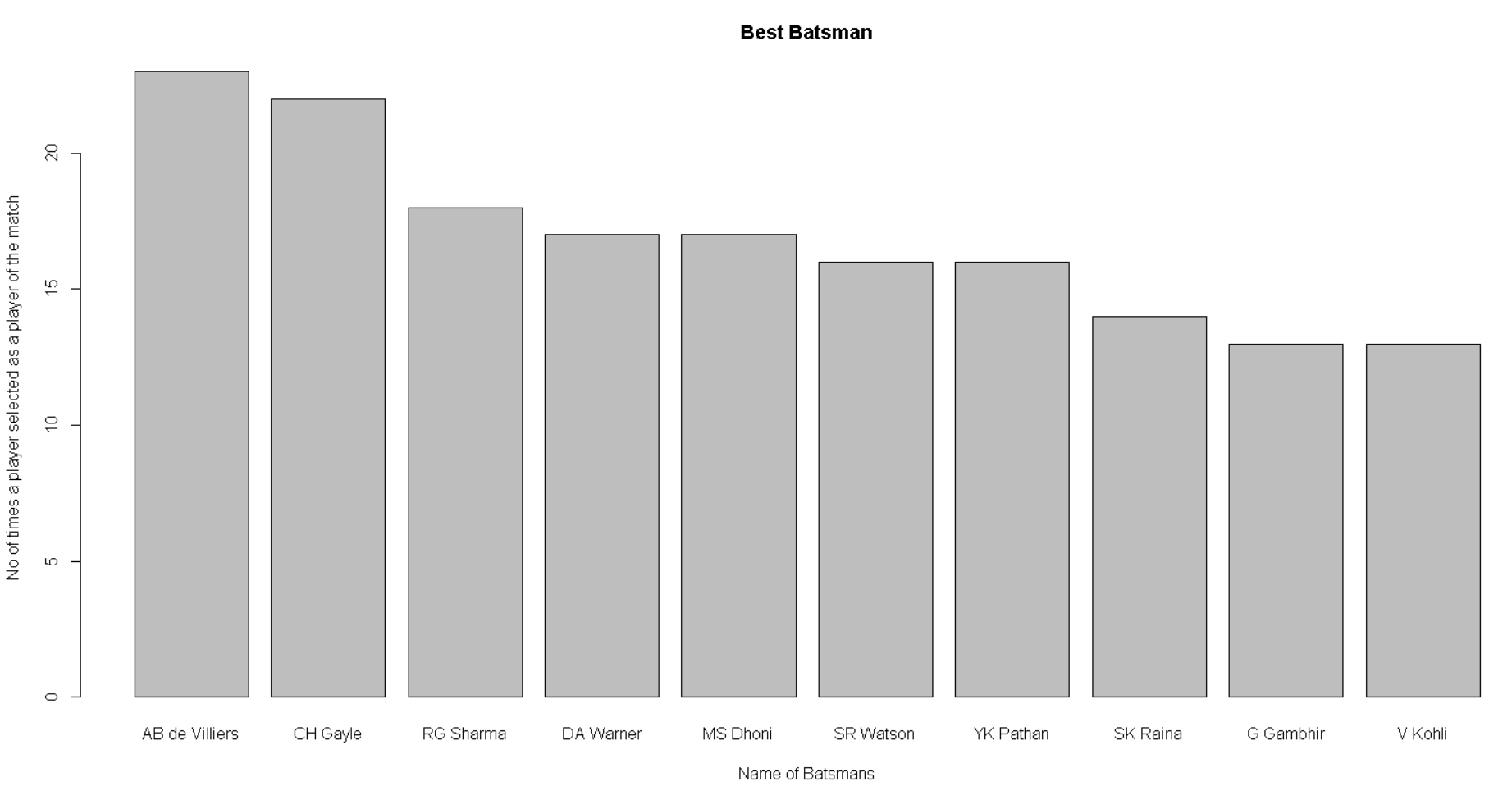
a<- table(match$player\_of\_match)

b<-sort(a,decreasing = TRUE)

c<- head(b,10)

c

barplot(c,main = "Best Batsman",xlab = "Name of Batsmans",ylab ="No of times a player selected as a player of the match" )



* AB de Villiers won most “Man of the Match” prize in IPL history, then Chris Gayle, later Rohit Sharma

**Best Bowler**

s<-balls%>%

group\_by(bowler) %>%

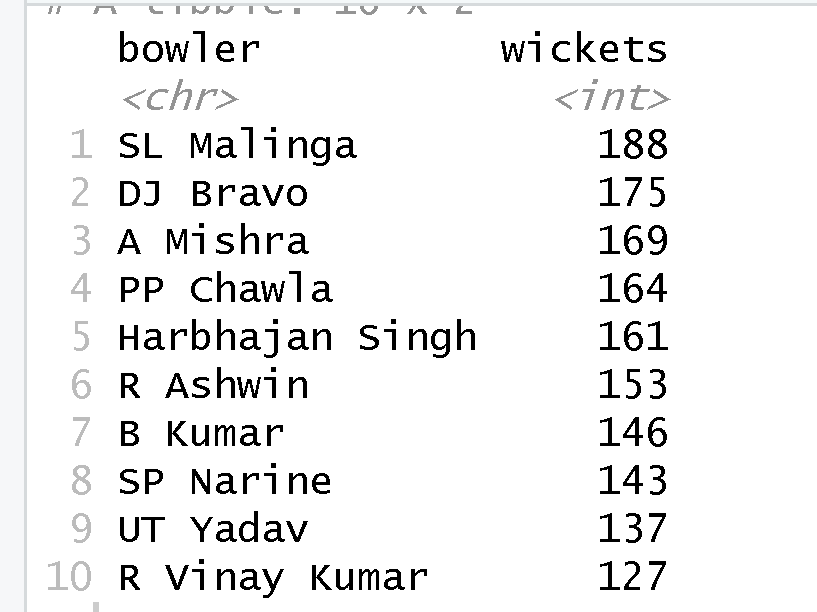
filter(is\_wicket==1)%>%

summarise(wickets=n())%>%arrange(desc(wickets))

s

t<-head(s,10)

t



* The player who took most wickets in IPL, is SL Malinga. Then, DJ Bravo, later Amit Mishra.

**Batsman with most** **No of Sixes**

#Top 10 batsman with most number of 6s in IPL

m<-balls%>%

group\_by(batsman)%>%

filter(batsman\_runs=="6")%>%

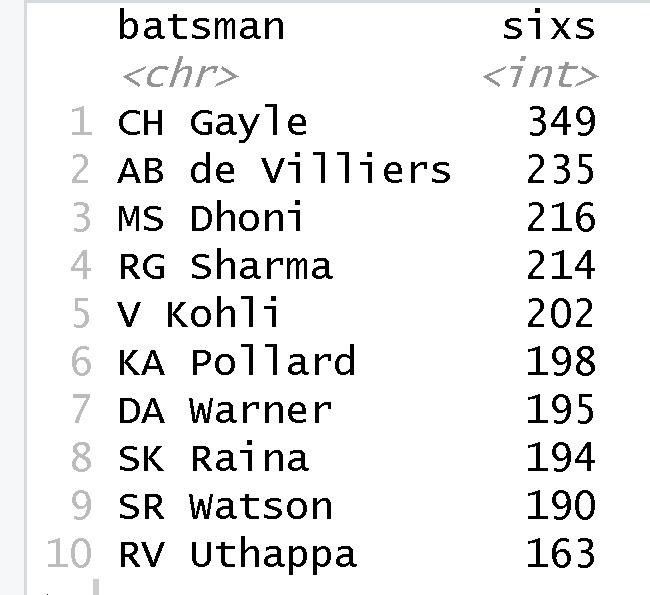
summarise(sixs=n())

m

n<-arrange(m,desc(sixs))

o<-head(n,10)

o



* The batsman who hit most sixes is Chris Gayle. Then AB de Villiers, later MS Dhoni.

**Which team is dominating in a certain location**

home\_ground\_advantage\_1<-match%>%

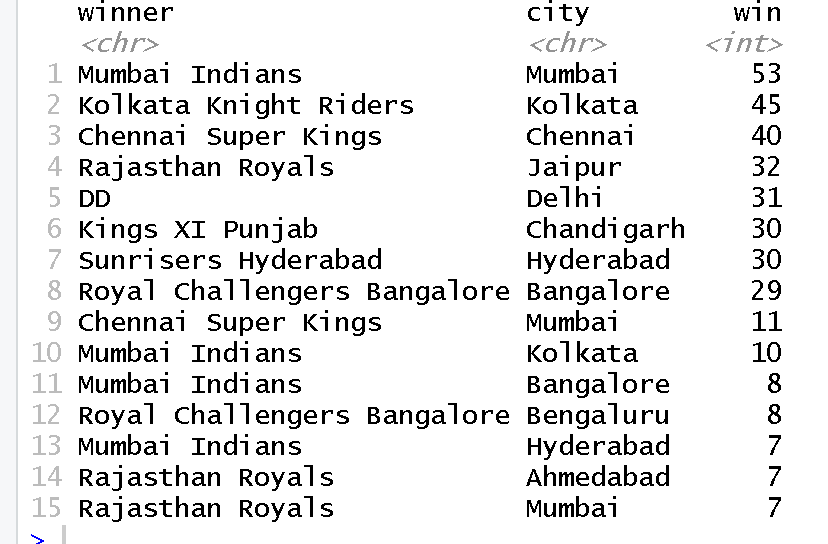
filter(result != 'no result') %>% group\_by(winner,city) %>%

summarise(win = n()) %>% arrange(desc(win))

View(home\_ground\_advantage\_1)

home\_ground\_advantage<- head(home\_ground\_advantage\_1,15)

home\_ground\_advantage



* Mumbai Indians won in Mumbai 53 times.
* Kolkata Knight Riders won in Kolkata 45 times.
* Chennai Super Kings won in Chennai 40 times.
* Rajasthan Royals won in Jaipur 32 times.
* Delhi Darevils (Delhi Capitals) won in Delhi 31 times.

**Batsman With most Runs**

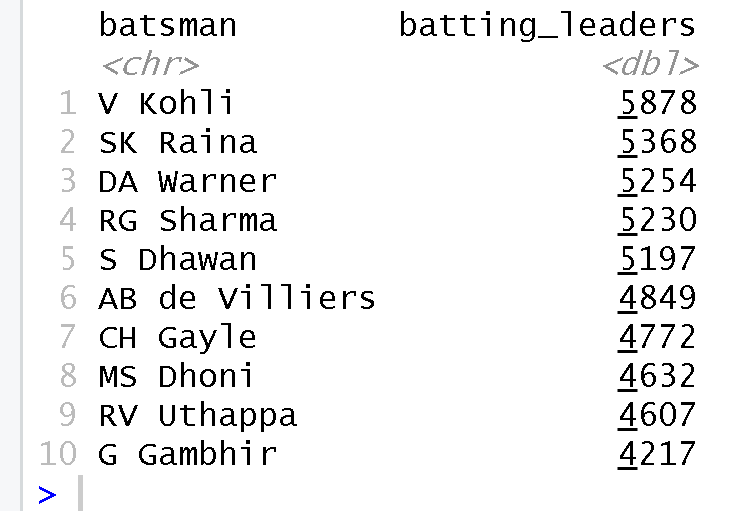
batting\_leaders<- summarise(group\_by(balls,batsman),batting\_leaders=sum(batsman\_runs))

batting\_leaders

batting\_leaders1<-arrange(batting\_leaders,desc(batting\_leaders))

top\_ten\_batsmans<-head(batting\_leaders1,10)

top\_ten\_batsmans



* Virat Kohli has scored most in IPL history.
* Second most run scorer is Suresh Raina.
* Third most run scorer is David Warner.

**What is a winning score when batting first?**

Here, we can see the association between proportion of matches won and total run scored by the team, batting first.

Here, the total run is given in column “runs\_tot” which is a numerical value and the value of matches won which is given in the column “Result” denoted as 0 and 1 which used as a categorical value. ( DICHOTOMOUS VARIABLE)

As the “Result” values are categorical in nature i.e., we cannot use linear regression hence we have used logistic regression in this model.

We define “Result” as y and “runs\_tot” as x here.

We have to get the value of log [ p / (1-p) ] here.

Where, p / (1-p) is value of the odds of the events which value can be greater than 1.

Here, log [ p / (1-p) ] = 0 + 1 . x = y (let)

Calculating we get, p = e^y / ( 1 + e^y ) , where p is the probability value which lies between 0 and 1.

And as there are only two categories as 0 and 1 which defines no failure or atleast 1 failure, we have used “binomial” as the family of the model.

Here, we can see in the summary of the model that,

Deviance residuals:

1Q value is -1.0024 and 3Q value is 1.0708 which is not perfectly symmetric but almost symmetric.

Median value is -0.5345 which is not exactly 0, AND OVER HERE MIN. VALUE AND MAX. VALUE ARE ALMOST SAME BUT OPPOSITE IN SIGN ,

So, it will follow normal distribution but not perfectly.

Coefficients:

Let, 0, and 1 be the coefficients for the intercept and for x respectively.

SINCE IT IS BASED UPON ONLY ONE DISTRIBUTION THEREFORE IT IS FOLLOWINF T DISTRIBUTION HENCE THE T VALUE

Now, testing whether intercept has some influence to predict the logs of odds or not.

H0: 0 = 0 vs H1: 0 ≠ 0.

p-value for the intercept is less than 2e-16 which is less than 0.05. So, we reject H0.

So, we can say that, there is an influence of intercept to predict the logs of odds of the model.

Now, testing whether the x has some effect to predict the logs of odds or not.

H0: 1 = 0 vs H1: 1 ≠ 0.

p-value for the intercept is less than 2e-16 which is less than 0.05. So, we reject H0.

So, we can say that, there is influence of x variable to predict the logs of odds of the model.

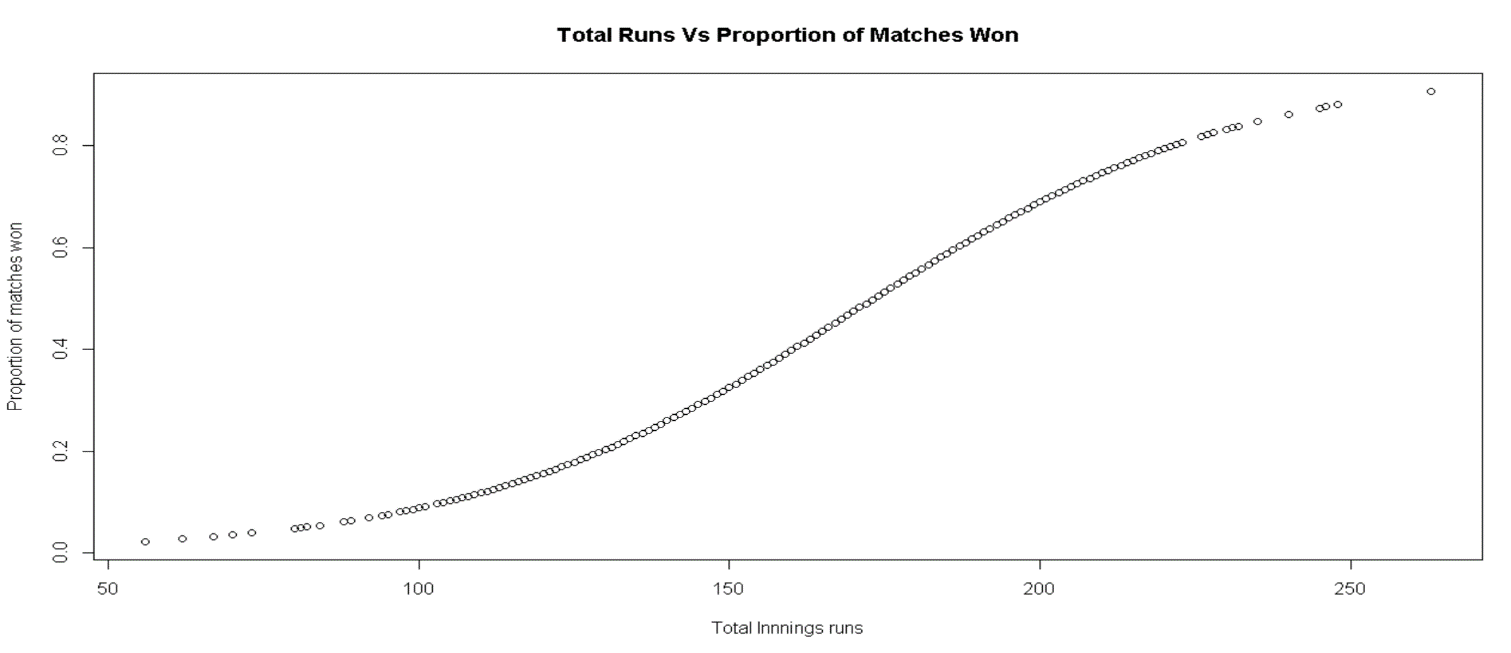
The AIC value here is 982.62 which is a relatively small value(the smaller the value, the better the model). So, we can say that, the fit is a good fit.

The number of Fisher Scoring iterations is 4. (the smaller the value, the better the model)

It is also a small value. So it IS PREDICTED AS A good fit and we might get the desired S-type graph for this logistic regression in our 4th try for this model. (For which the value of the product of all the probability is maximum { AFTER DOING THE MAXIMUM LIKELIHOOD}).

***Regression formula, based on the results of the summary coefficients :***

p <- exp(-5.581237 + 0.032513\*new\_data$runs\_tot)/(1 + exp(-5.531237 + 0.032513\*new\_data$runs\_tot))



The graph is showing the proportion of matches won and runs scored. Clearly, there is a very close association between scoring more runs and the probability of victory. Based on this analysis, a team has a 50% chance of winning if they score 170 runs in their inning, and an 80% chance of winning if they score 210.

CONCLUSION:

More detailed data is required to increase the accuracy of these models.

It was found that

a) teams are at a slight, but significant, disadvantage by batting first,

b) total number of runs was unsurprisingly the strongest predictor of match outcome,

c) scoring zero or one run from a ball decreases the probability of a win.

That being said, these results are still informative: Teams armed with the knowledge that scoring one run from a ball could theoretically place less emphasis on strike rotation (i.e., scoring one run frequently) and greater emphasis on having batsmen face bowlers in more favorable matchups, thereby likely increasing the number of runs scored in the long-term context of an inning, and being more likely to win the match.

**REFERENCES**

1. **Data Source :** https://www.kaggle.com/patrickb1912/ipl-complete-dataset-20082020
2. **Books:** (i) Applied Logistic Regression by David Hosmer, S. Lemeshow, R. Sturdivant

(ii) Practical guide to logistic regression by Joseph M.Hillbe

|  |
| --- |
|  |