S670 – EXPLORATORY DATA ANALYSIS FINAL PROJECT

Studying Toss and Home Advantage in IPL

Team Members – Meet V, Ojaas H, Vijeet S.

Cricket is one of the most widely followed sport in the world. Club cricket in India is a very popular limited overs version of the game – Indian Premier League (IPL). The IPL takes place every year from its inception in the year 2008 to 2019. IPL is usually played in the summer months of April and May across various Locations in India. There are 8 major teams in IPL, each of which have home ground in different cities across India.

The questions that we are trying to answer here are –

1. How runs and wickets vary per over across the seasons? Does no of runs scored have any relation with the no of wickets? Who is the highest run scorer each season? Who is the highest wicket-taker each season?
2. How does the toss winning and toss decision affect the outcome of the match in IPL? Is toss advantage a real thing in the IPL? Can we predict the chances of the team winning the match on winning the toss?
3. Does a team playing on the home ground have a significant advantage over the away team? Can we predict the chances of a team winning the match given that it plays at home ground?

All the above questions matter in the world of cricket. It is generally assumed that a team performs very well in their home ground. However, would that be the case for all the teams that participate in the IPL? Also how does winning the toss and the subsequent toss decision have any effect on the game alongside the home advantage can be useful to predict which team has the best chances to win the tournament in this season.

Whenever the chances of rain are forecasted during a match, we can safely assume that Duckworth-Lewis-Stern Method maybe applied in the match. DLS method is used when the game is interrupted due to rain and the targets for the second team are adjusted according to the resources available to the second team to create proper revised targets. DLS method considers the factors like runs scored and wickets lost by first team, similar statistics available for the second team, and a few other factors to set the revised score. Knowing the toss advantage and home advantage will help us in such a case, because we can then determine the possibility of a team winning with revised targets that have been set. In case of rains interrupting a match, we can still make a valid prediction as to how the teams fare with respect to the odds of winning.

Apart from the above said advantages, we the team members have been avid followers of cricket for a long time now and each of us has a favorite team and we always have healthy debates about which of the teams playing has the best chances to win the game. This model would also help us to better understand how the game of cricket works at a competitive level. In this project, we got to learn a lot about how the statistical modelling can be applied to sports to gain valuable insights.

DATA SET DESCRIPTION

Matches.csv includes –

- Id - Numeric Id of every match which has been played.
- Season - Year in which a match has been played.
- City - City in which match is played.
- Team 1 - Name of the team who played the match.
- Team 2 - Name of another team who played the match.
- Toss_winner - Name of the team who wins the toss.
- Toss_decision - Team who wins the toss get a chance to decide whether to bat or field. They have two options – Bat or Field.
- Winner - Name of the team who wins the match.
- Win_by_runs - Win by Runs is when winning team bats first and they win by certain number of runs.
- Win_by_wickets - A team can win a match by a certain number of wickets.
  This means that they were batting last and reached the winning target with a certain number of batsmen still not dismissed.

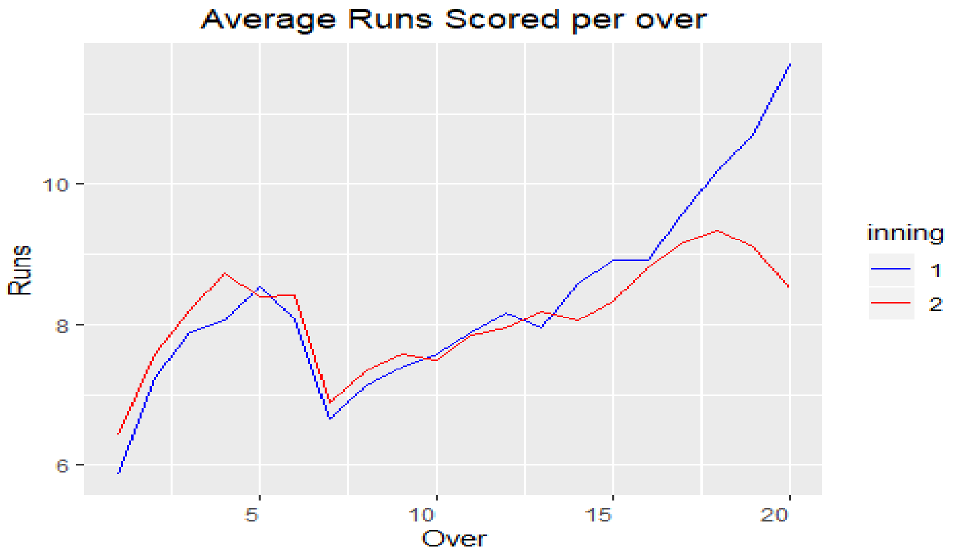
Deliveries.csv includes –

- Match_id - Numeric Id of every match which has been played.
- Inning - One of the divisions of a match during which one team bowls and another team bats and vice versa.
- Batting_team – Name of the team who is batting.
- Bowling_team - Name of the team who is bowling.
- Total_runs – Total runs scored in each ball.
- Batsmen – Name of the Batsmen who is batting.
- Bowler - Name of the Bowler who is Bowling.
- Over – An over consists of six consecutive legal deliveries bowled from one end of a cricket pitch to the player batting at the other end, almost always by a single bowler.
- Ball – A Legal Delivery bowled by a bowler.
- Player_dismissed – A dismissal occurs when a batsman's period of batting is ended by the opposing team. This feature will have the name of the batsmen being dismissed by the bowler.
- Dismissal_kind – Ways in which batsmen can be dismissed in the cricket. It contains values like caught, bowled, runout, caught and bowled, etc.

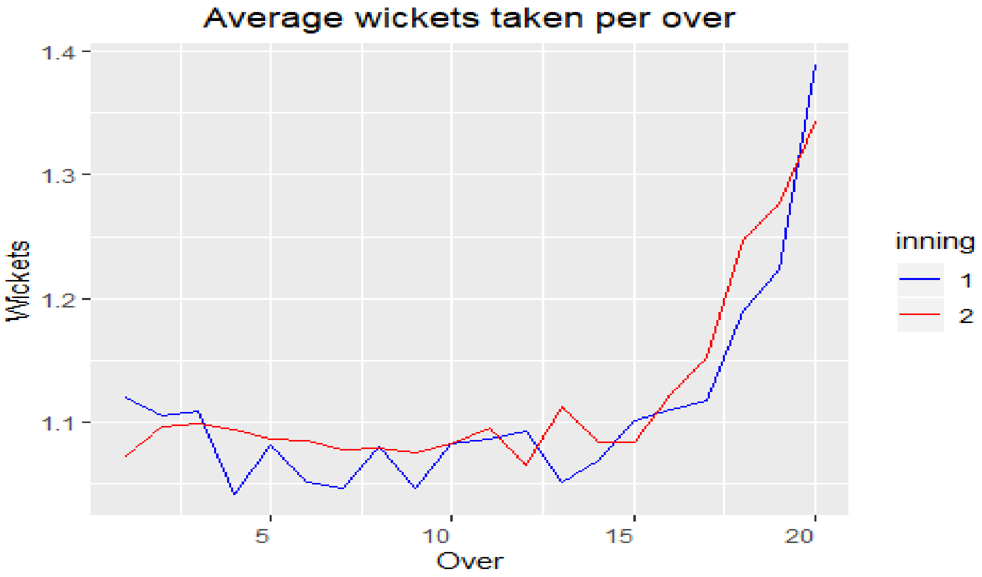
Hometeams.csv includes –

- City – City in which home ground of a team is situated.
- Hometeam – Name of the team.
- Abbr – Abbreviatoin of the home team.

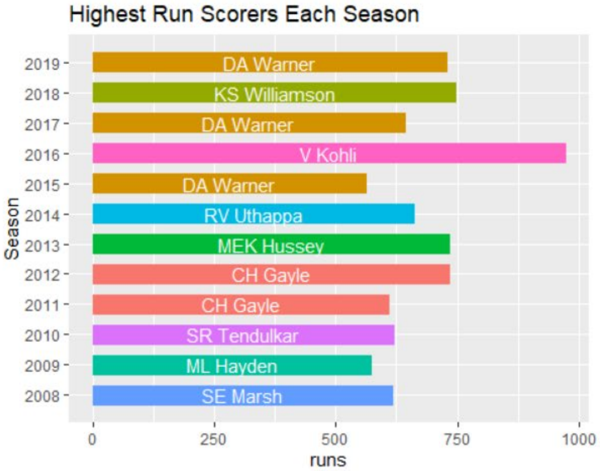
PART 1: FINDING THE INHERENT TRENDS IN DATA

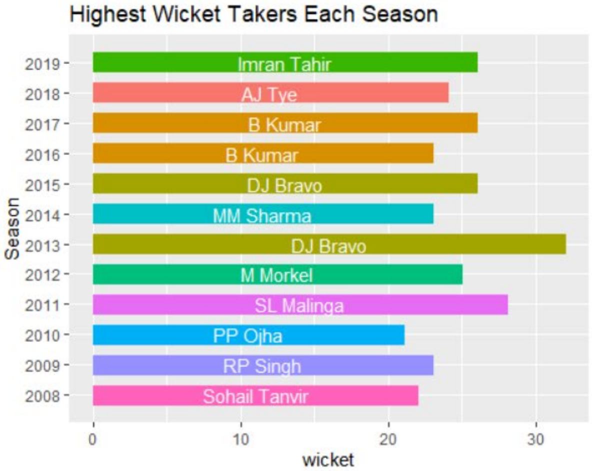


Initial few overs in every innings of the IPL are considered are considered as batting powerplay. In batting powerplay, it is mandatory for the fielding team to keep two players outside the 30 yards circle and all the other players inside the circle. This provides a significant batting advantage to the teams as there are only two catching hotspots outside the 30 yards circle, hence it makes hitting boundaries easier. We can see from the graph that the average run rate is high for the first few overs and then slumps a little and then picks up again towards the end of the inning. The end of the innings i.e. overs 16-20 are considered death overs in T20 cricket. These overs are the last few overs of batting for any team and the goal of any batting team is maximize the score as much as possible. This implies that the batting team does not usually bother about the wickets in this phase of the match and just focuses on hitting boundaries. Hence, we can see that the average runs per over is very high towards the end of the innings.



It is a known fact in the IPL that the death overs are used by the batting team to hit boundaries and catch outs are the most common type of dismissal in the IPL. The trend shows that the average number of wickets lost in the death overs is higher than the average wickets lost in the initial overs across all the seasons of IPL.

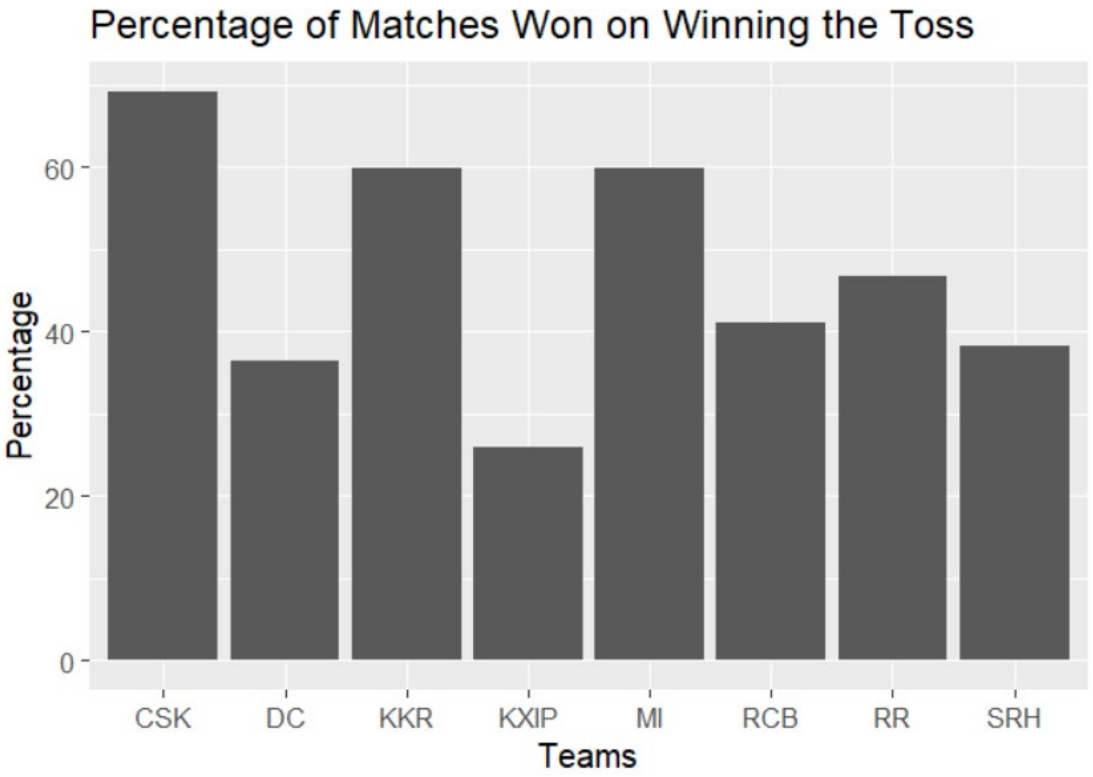




This graph shows the highest scorer in each season of the IPL since its inception. IPL has a format of player bidding in which the participating teams can bid for players. The base prices for the players are determined by the corresponding performance metrics in the previous season. This implies that the highest scorer players from the previous seasons always have a higher bidding base price. From the graph above, we can make an estimate that the IPL season 2016 was highly supportive towards batting as compared to all other seasons. The highest scorer in each season is awarded the Orange cap as a symbolic gesture apart from the monetary benefits.

Like we took the highest run scorer to determine the performance metrics for the batsmen, in a similar way, the number of wickets taken is used as the performance measure quantity for the bowlers. The bowler who gets the highest wickets in season sees a corresponding increase in the bidding prices next season. Just like the Orange cap is awarded to highest scoring batsman, the highest wicket taking bowler is awarded a Purple cap.

PART 2: IS TOSS ADVANTAGE A REAL THING IN IPL?

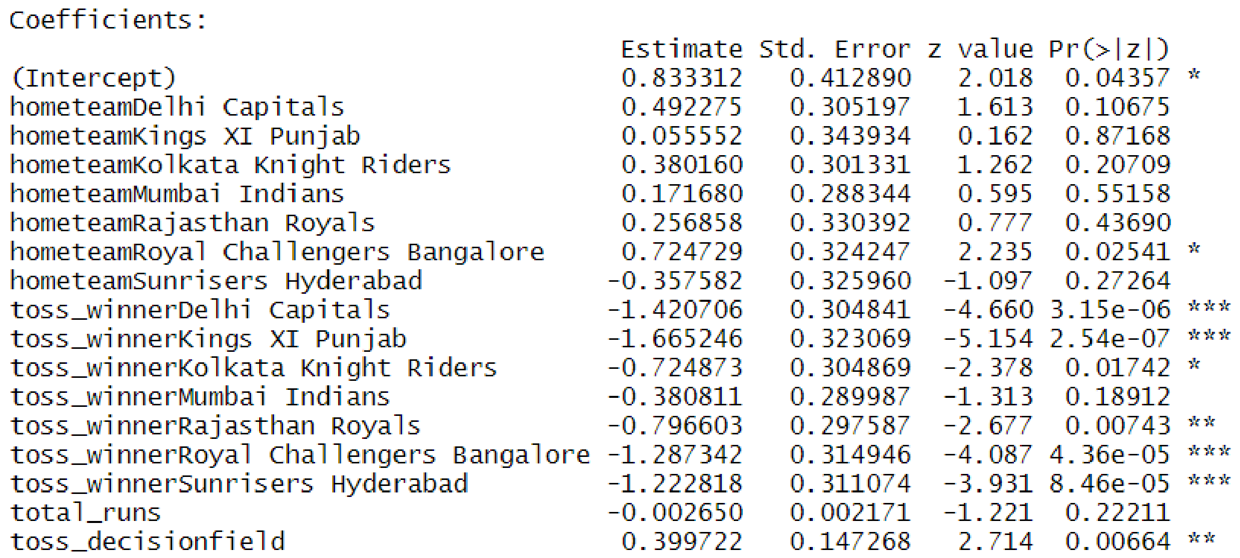


The graph above shows the results that we found on studying the existing data. We can see very clearly that the probability of winning the game on winning the toss varies highly among the teams. If we consider some of the teams like Chennai Super Kings, Kolkata Knight Riders and Mumbai Indians, we can see that the probability is higher than 50% that they win the match on winning the toss. Also, when we have a look at teams like Kings XI Punjab and Delhi Capitals, we can see that the probability to win even after winning the toss is less. This leads us to the question that whether any concept of toss advantage does really exist in the IPL? To answer this question, we decided to build a predictive analysis model.

We have fitted a Gaussian Linear Model (GLM) to the data. A GLM is a more general version of a linear model: the linear model is a special case of a Gaussian GLM with the identity link. We fit GLMs because they answer a specific question that we are interested in. The slope and the intercept of such a curve account for a tendency of extreme risk to tend toward 0/1 probability. The feature that we are predicting here is whether the team wins the toss also wins the match. The variables that are used to predict the target variable are Home Team, Toss Winner, Toss Decision and Total Runs. The model takes the effect of all these variables into account to predict the probability of the team that wins the toss also wins the match.

Model Definition –

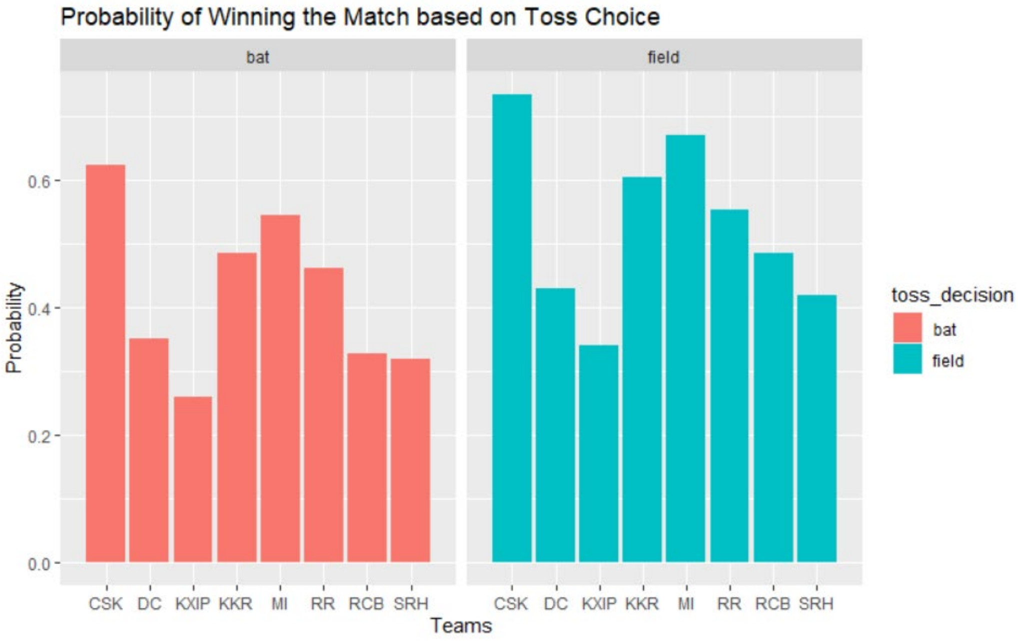
```
glm(formula = tossAdvantage ~ hometeam + toss_winner + total_runs +
    toss_decision, family = binomial, data = modelData)
```



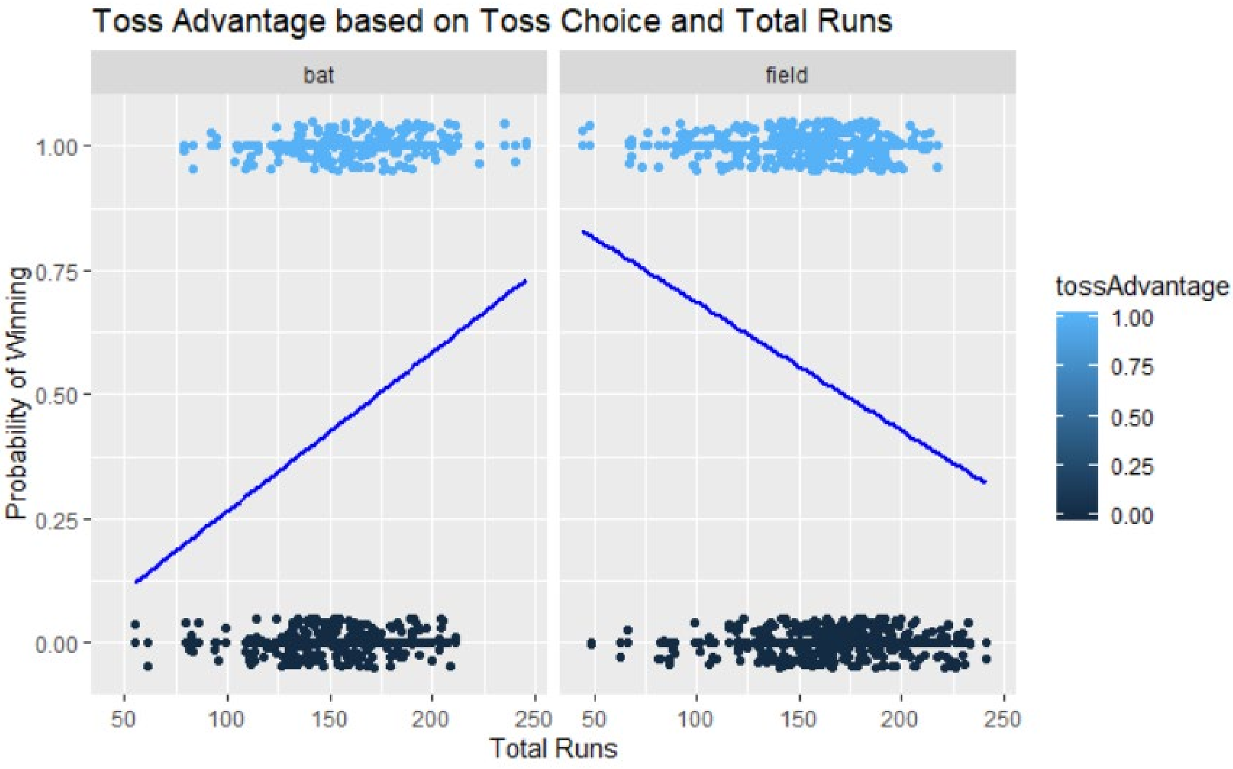
```
Coefficients:
                                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)                              0.833312   0.412890   2.018  0.04357 *  
hometeamDelhi Capitals                   0.492275   0.305197   1.613  0.10675    
hometeamKings XI Punjab                  0.055552   0.343934   0.162  0.87168    
hometeamKolkata Knight Riders            0.380160   0.301331   1.262  0.20709    
hometeamMumbai Indians                   0.171680   0.288344   0.595  0.55158    
hometeamRajasthan Royals                 0.256858   0.330392   0.777  0.43690    
hometeamRoyal Challengers Bangalore      0.724729   0.324247   2.235  0.02541 *  
hometeamSunrisers Hyderabad             -0.357582   0.325960  -1.097  0.27264    
toss_winnerDelhi Capitals               -1.420706   0.304841  -4.660 3.15e-06 ***
toss_winnerKings XI Punjab              -1.665246   0.323069  -5.154 2.54e-07 ***
toss_winnerKolkata Knight Riders        -0.724873   0.304869  -2.378  0.01742 *  
toss_winnerMumbai Indians               -0.380811   0.289987  -1.313  0.18912    
toss_winnerRajasthan Royals             -0.796603   0.297587  -2.677  0.00743 ** 
toss_winnerRoyal Challengers Bangalore -1.287342   0.314946  -4.087 4.36e-05 ***
toss_winnerSunrisers Hyderabad          -1.222818   0.311074  -3.931 8.46e-05 ***
total_runs                              -0.002650   0.002171  -1.221  0.22211    
toss_decisionfield                       0.399722   0.147268   2.714  0.00664 ** 
```

By observing the coefficients for the toss_winner feature, we can see that, with respect to Chennai Super Kings, Kings XI Punjab lose 16.6% more matches after winning the toss, on the hand, Mumbai Indians just lose about 3.8% more matches in comparison. Also, from the model coefficient's above, we can see that in case of toss decision to field, the teams tend to win 3.9% more than when they opt to bat first after winning the toss.

We can have a look at the ROC curve to see the fit of the model on the data. The area under the curve determines the amount of data that is fit by the regression model. In our case, the area under the curve comes out to be 0.6364. The accuracy of the prediction data also comes out to be 63.62%. When we look at the coefficient of the fits for various factors involved in the model, we can see that the toss decision plays an important role in the determination of the which team goes on to win the match.



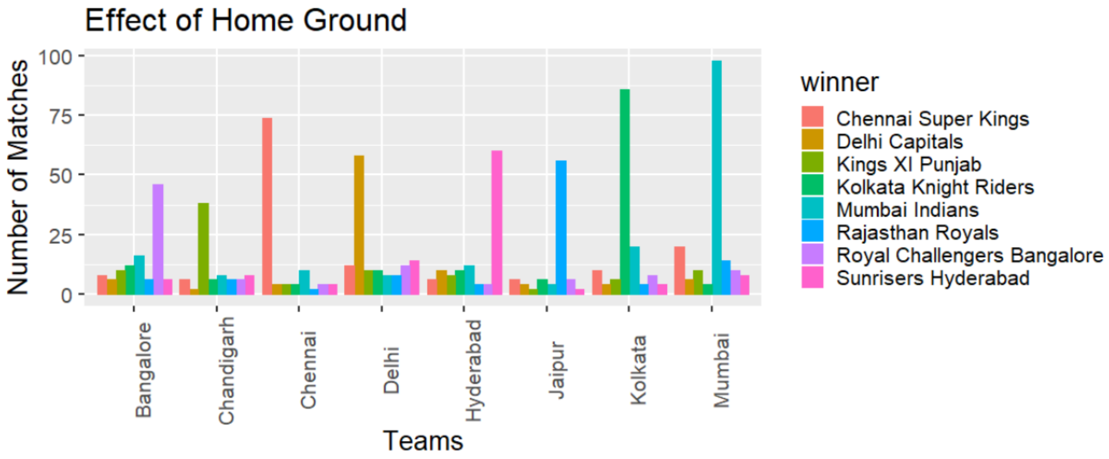
The graph above shows the effect of the toss decision on winning the game provided that the team wins the toss. We can see that overall, the teams that win the toss and chose to bat have lower chances of winning the match than the teams that win the toss and opt to field. This proves our intuition that the toss decision apart from winning the toss affects the outcome of the match in some ways.



We decided to have a look at one other factor that can affect the outcome of the match in a very serious way apart from the toss outcome and toss decision. We see that the team that bats first has a high chance of winning if they manage to score above 200 runs. And the trend is exactly the reverse for the fielding first. If a team has chosen to field first, they must restrict the batting team to a score below 200 if the fielding team wants to have a fair shot at winning the game.

From this analysis, we can say that toss advantage exists in the IPL, but a very few teams manage to make use of it wisely.

PART 3: IS HOME ADVANTAGE A REAL THING IN IPL?

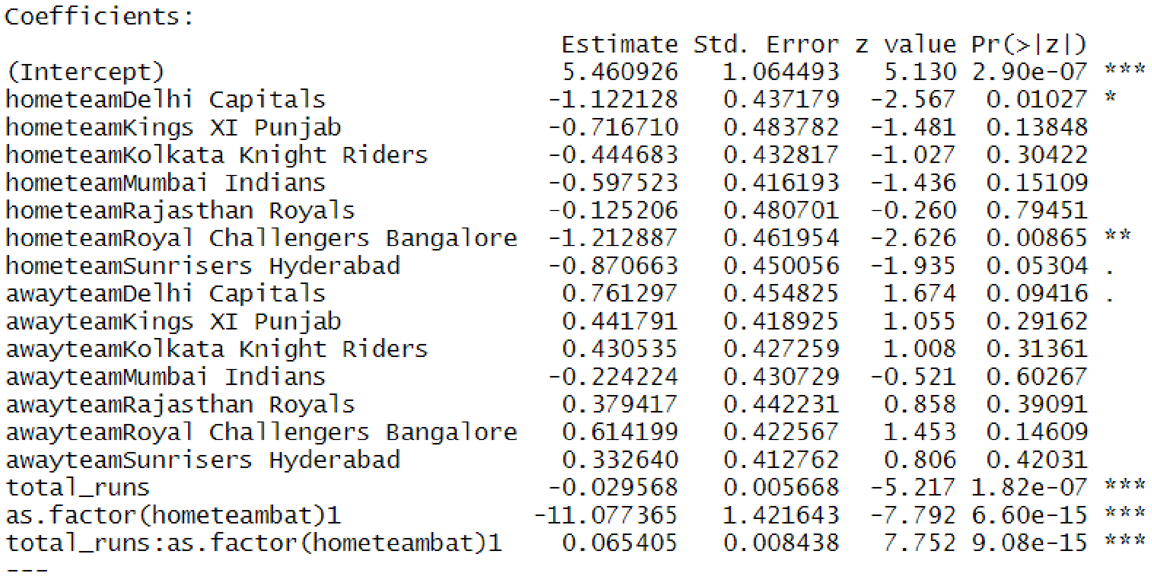


On studying the data that we have, we found out that the teams win a lot of matches at their home grounds as compared to away locations. This is a very logical conclusion, given that the format of the IPL states that the team must play half of its matches in playoffs at its home location. This fact is also confirmed by the graph above. To find out if home advantage is a real thing or not, we decided to build a statistical model to predict if the home team can win the game or not, and if so by what percentage.

The model used here is again a Gaussian Linear Model (GLM), same as the one used in toss advantage question above. To predict the chances of team winning a match at their home ground, we decided to use the predictor variables of Home team, Away team, total runs and whether the home team bats first or not.

Model Definition –

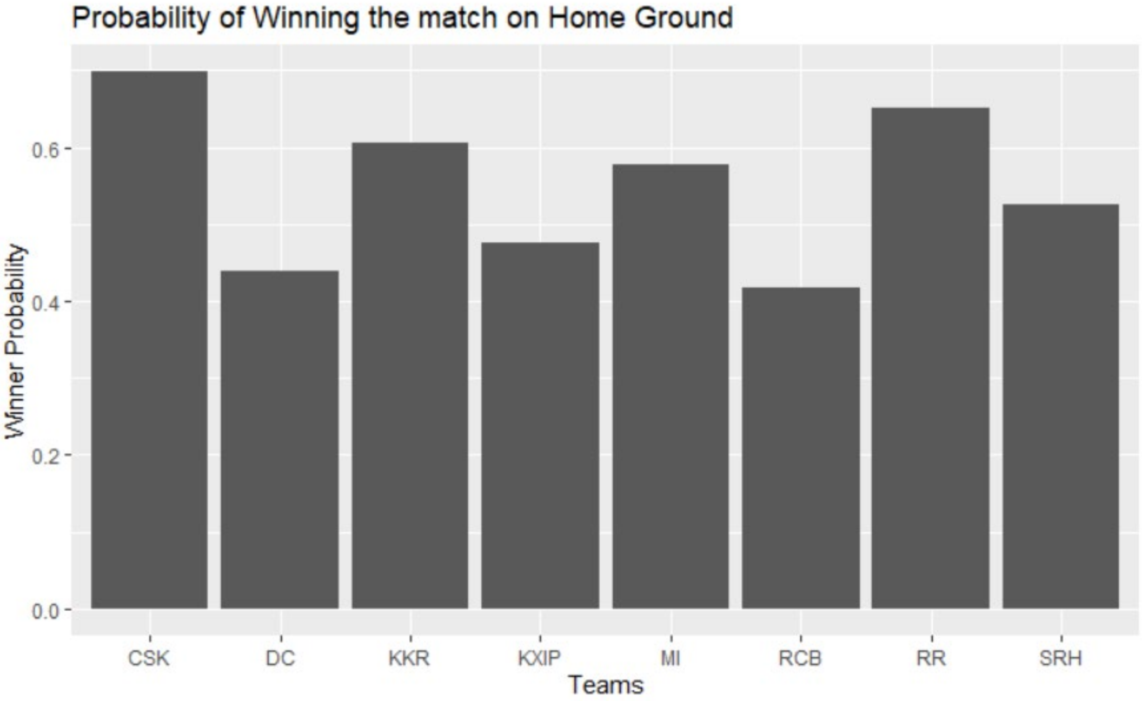
glm(formula = homewinner ~ hometeam + awayteam + total_runs *
as.factor(hometeambat), family = binomial, data = home.adv)



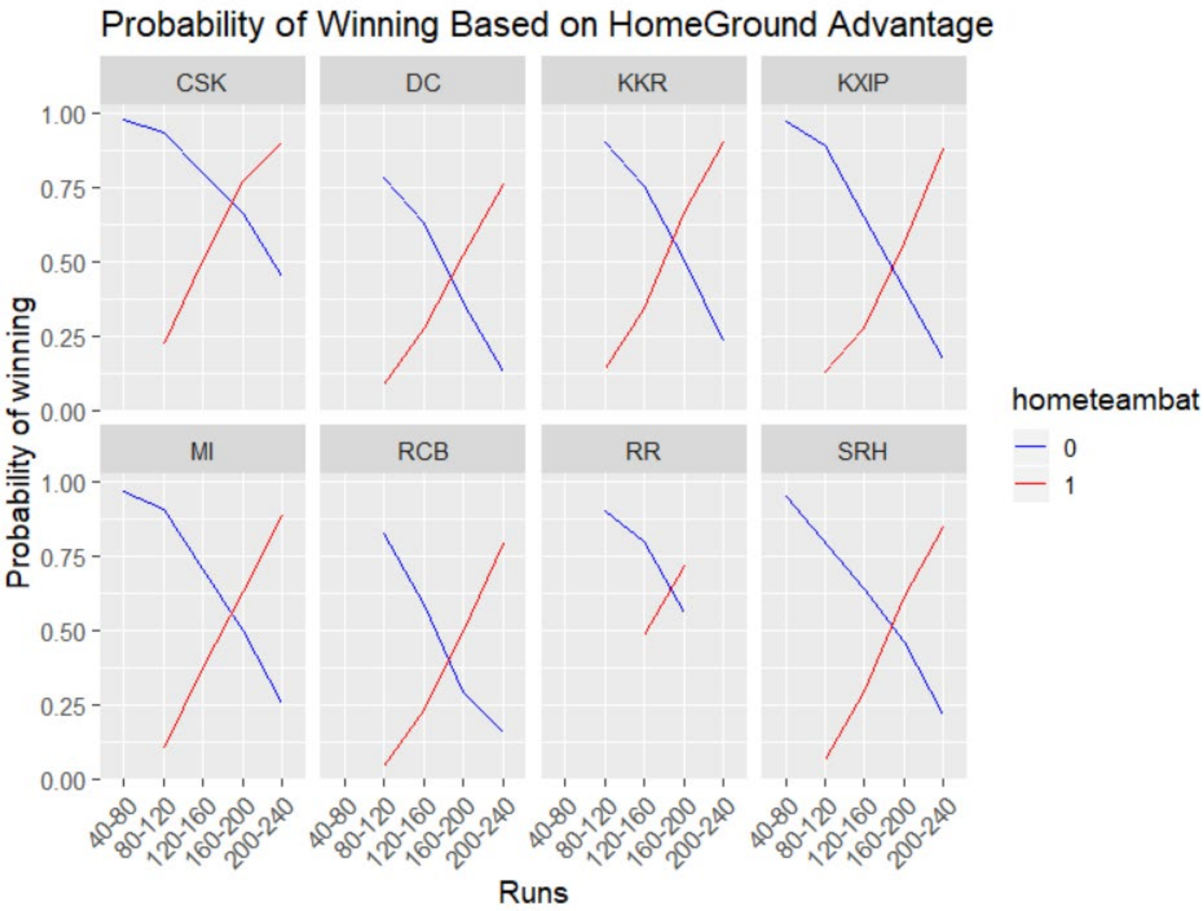
```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)                       5.460926   1.064493   5.130 2.90e-07 ***
hometeamDelhi Capitals           -1.122128   0.437179  -2.567  0.01027 *  
hometeamKings XI Punjab          -0.716710   0.483782  -1.481  0.13848    
hometeamKolkata Knight Riders    -0.444683   0.432817  -1.027  0.30422    
hometeamMumbai Indians           -0.597523   0.416193  -1.436  0.15109    
hometeamRajasthan Royals         -0.125206   0.480701  -0.260  0.79451    
hometeamRoyal Challengers Bangalore -1.212887   0.461954  -2.626  0.00865 ** 
hometeamSunrisers Hyderabad      -0.870663   0.450056  -1.935  0.05304 .  
awayteamDelhi Capitals            0.761297   0.454825   1.674  0.09416 .  
awayteamKings XI Punjab           0.441791   0.418925   1.055  0.29162    
awayteamKolkata Knight Riders     0.430535   0.427259   1.008  0.31361    
awayteamMumbai Indians           -0.224224   0.430729  -0.521  0.60267    
awayteamRajasthan Royals          0.379417   0.442231   0.858  0.39091    
awayteamRoyal Challengers Bangalore  0.614199   0.422567   1.453  0.14609    
awayteamSunrisers Hyderabad       0.332640   0.412762   0.806  0.42031    
total_runs                       -0.029568   0.005668  -5.217 1.82e-07 ***
as.factor(hometeambat)1         -11.077365   1.421643  -7.792 6.60e-15 ***
total_runs:as.factor(hometeambat)1  0.065405   0.008438   7.752 9.08e-15 ***
---
```

The coefficients “hometeam” are relative to Chennai Super Kings. The coefficient for Delhi Capital implies that compared to CSK they lose 11% of the match at home grounds while Rajasthan Royals lose only 1.2% of the matches played at home. Similarly, for “awayteam” coefficients suggest that compared to CSK, Delhi Capitals win 7.6% of their away matches. So, the model suggests teams like CSK, RR, MI and KKR grab the home advantage whereas other teams are not so efficient at capturing on the chance of Home Advantage. The option to bat first also plays an important role while playing at Home. It is preferred to field first at home. The negative coefficient “hometeambat” concludes this.

The ROC curve above has an area of 0.7052, which implies that the model fits about 70% of the data properly. When we calculate the accuracy of the predictions given by the model, we can see that the accuracy is 70.64%, which means that the model can predict the data well.



When we have a look at the graph for the chances of the team winning the game provided that they are playing on their home ground, we can see that the chances of winning are higher for some teams like Chennai Super Kings, Rajasthan Royals and Kolkata Knight Riders, while some teams like Royal Challengers Bangalore and Delhi Capitals fail to capitalize on the home advantage in the matches.



The above graph shows how the batting order and the number of runs scored affect the chances of the team winning a game at their home. 0 in the hometeambat implies that the home team fields first, while 1 indicates that the home team bats first. We can see very clearly for every team that the chances of winning for the home team increase when they chose to bat and score runs greater than 160, and a corresponding result is also derived when the home team fields first. We can see that the home team must concede runs less than 160 if they want to have a safe chance to win. If the home team choses to field first and leaks more than 160 runs, their chances of winning the game are affected highly. From the model predictions and the graphs above, we can state that Home Advantage exists in IPL for some teams and some other teams fail to capitalize on it.

CONCLUSION –

From the above exploratory analysis, we can say that there is a certain level of toss winning advantage that exist sin the IPL. Not all the teams that participate in the tournament are able to capitalize on the toss advantage factor, but it does exist to a great extent in the tournament. As for the home ground advantage, some teams tend to win more on their home ground as compared to other teams. Overall, we can say that the concept of Toss Advantage and Home Advantage does exist in the IPL but is applicable only to a subset of teams that participate in the games.

LIMITATION –

The dataset that we have covers a lot of basic stuff about the game of cricket, but there are a lot more factors that are involved in the game apart from the names of the teams, venue, toss winner, toss decision , number of runs scored and number of wickets lost. Some of the other important factors that affect the game of cricket are the pitch conditions, weather conditions, team overall ratings, individual player ratings, form of the player etc. Knowing all of the above variable, or a subset of them would help us to provide a better model to predict the chances of a team winning a game provided that they win the toss and the chances of a team winning on the home ground.

FUTURE WORK –

If the information regarding the pitch type and conditions are available, we can make of estimate of how the chances of winning are affected with respect to batting order. If the weather condition forecasts rain, we know that Duckworth Lewis Method can be applied in the game. In such a case, the target revision is done based on various resources, hence the chances of winning the game is affected highly. If the weather is available as a feature in the dataset, then the chances of winning can be calculated accounting for the possibility of a DLS method as well as the performance of the teams during certain weather conditions and pitch types.

Apart from these, the overall ratings of the teams and the form of the players would help us to determine why some teams are able to capitalize on toss and home advantage, while some other teams cannot cope up with the advantage.