Visualization Project

Aniket Tiwari

2023-09-29

Introduction

In the fast-paced world of modern cricket, the **Indian Premier League (IPL)** stands as a shining beacon of sporting entertainment, blending talent from around the globe in a high-octane cricketing extravaganza. The IPL has not only redefined the way cricket is played but has also created an unprecedented wealth of data, ranging from player statistics to match outcomes, that provides a treasure trove of insights waiting to be unearthed. In this data analytics and *Visualization project*, we embark on a journey to delve deep into the IPL data-set(2008-2019), employing data analysis techniques and *powerful visualization* using R to unlock hidden patterns, uncover meaningful trends, and gain a comprehensive understanding of the league's evolution over the years.
Join us as we explore the numbers behind the IPL, shedding light on the strategies, performances, and stories that have shaped this iconic cricketing phenomenon.

Data Description

Data set Name: IPL Datasets

Source: Kaggle - IPL Datasets on Kaggle

Description: The IPL Datasets is a collection on Kaggle includes multiple CSV files, each focusing on specific aspects of the IPL out of which some of them used here are : 1-Overs Without Super Overs 2-Matches 3-Overs_Data

'Matches': This dataset provides detailed information about each IPL match, including match date, venues, teams, toss details, and more.

'Overs Without Super Over' : This data sets provides information regarding current innings , match id, batting team, bowling teams, runs conceded etc.

'Overs_Data' : This data set contains info about the runs made by each team after each match.

Key Features:

Detailed match information, allowing for analysis of match outcomes, venues, and team performance. Overs data contains over by over conceded runs which can be be very useful to study several aspects such as good rotation of bowlers by respective 'captains' or teams performance in *Power plays(01-04)* and *Death Overs(16-20)*.

Data Size: The dataset comprises multiple CSV files, with varying sizes depending on the specific dataset. For example, the "Matches" dataset contains 757 rows and 17 columns.

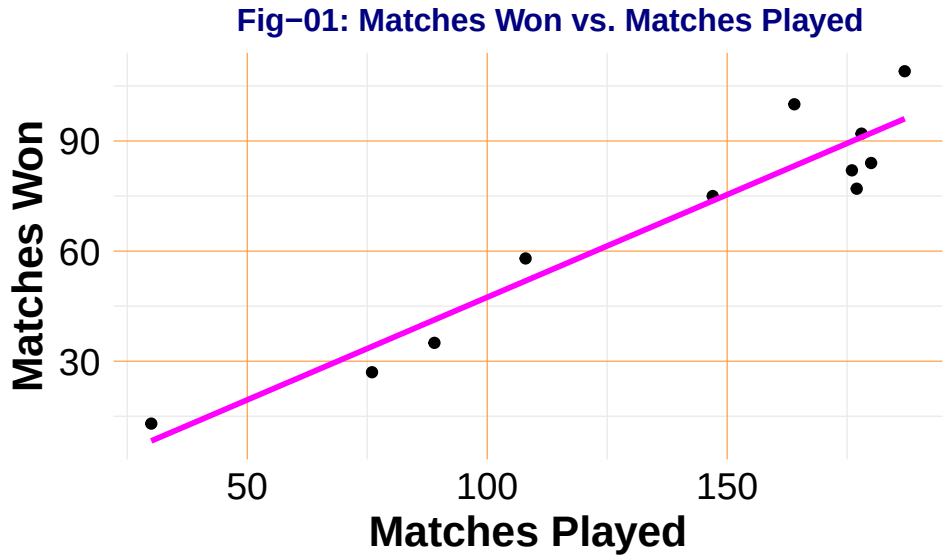
Data Format: All datasets are provided in CSV format, making them compatible with a wide range of data analysis and visualization tools.

Data Variables:

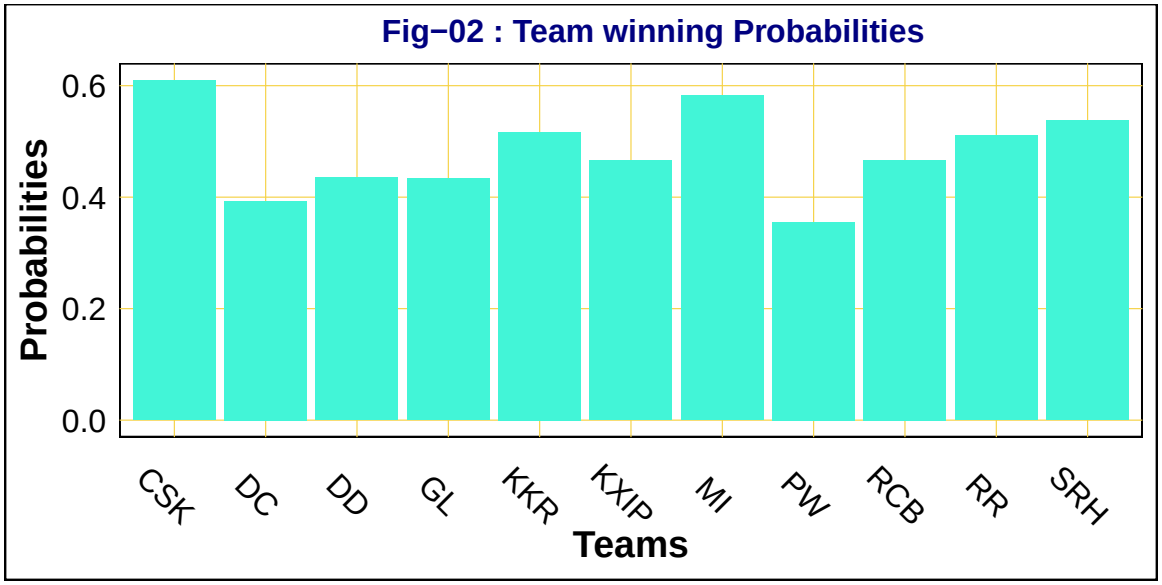
Matches : match_id, season, date, venue, team1, team2, toss_winner, toss_decision, result, dl_applied, winner, win_by_runs, winning team, win_by_wickets.

Overs Without Super Over : match_id, inning, batting_team, bowling_team, over, runs, wickets winner.

Exploratory Data Analysis

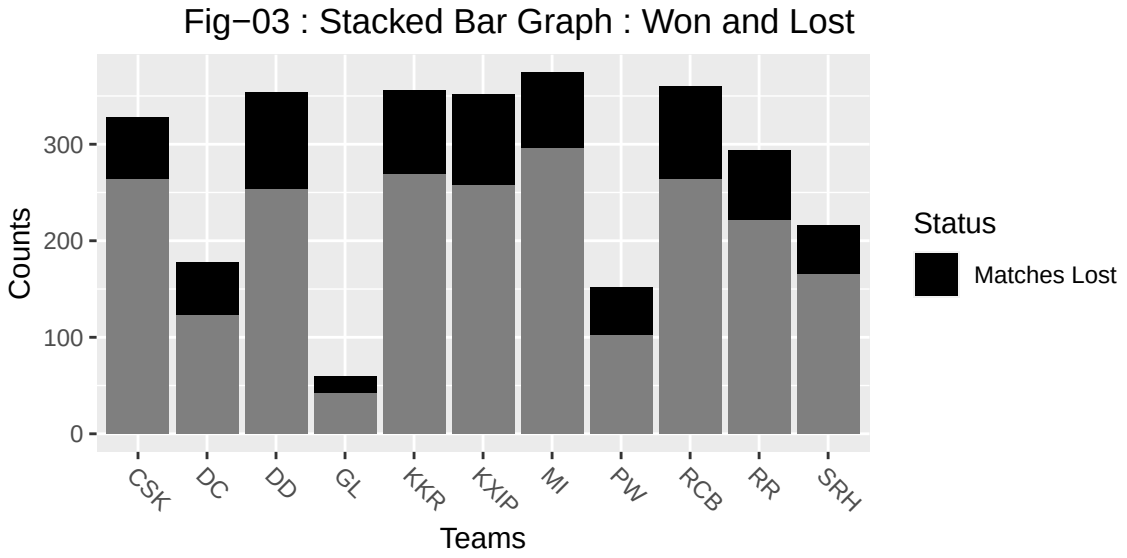


Above scatter plot (fig-01) represents a points plotted on the basis of 'matches won by teams' against the 'matches they played'.This is evaluated as the sum of each teams no. of matches they participated in and sum of no. of matches which they won.The blue line represents the 'best fit line' of the given data.

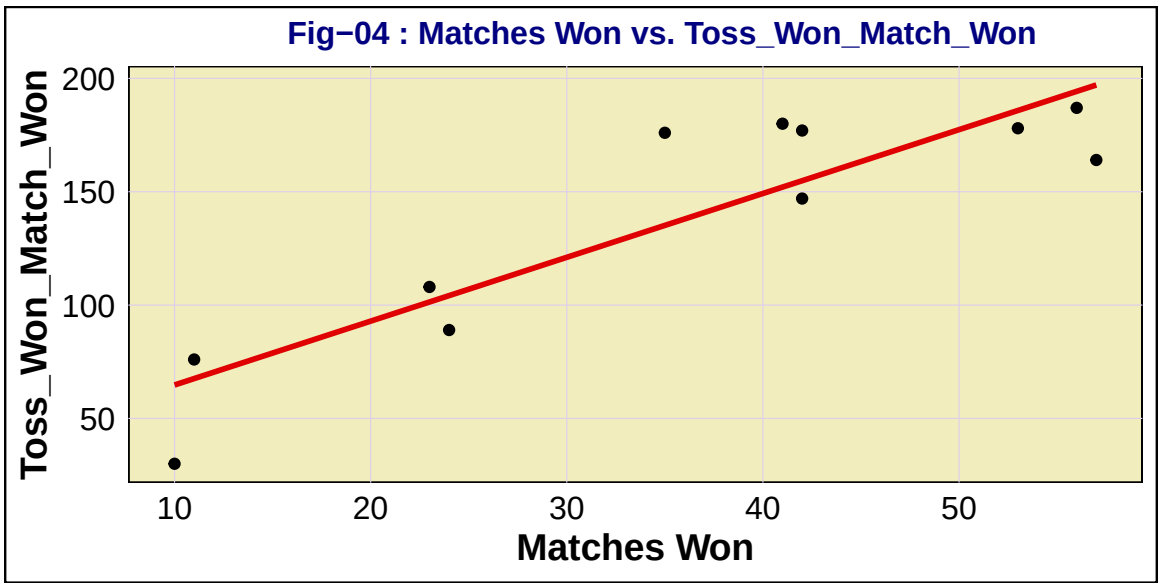


Above bar graph(fig-02) demonstrates the teams performance and its winning probabilities according to the previous performance.

From here it is clearly evident that teams **Chennai Super Kings(CSK)** and **Mumbai Indians(MI)** have the *highest* match winning probabilities i.e 0.62 and 0.57 respectively which means that CSK wins almost 60% of the times they play whereas MI wins almost 57% times.
Furthermore, "Pune Warriors(PW)" has the least percentage of winning i.e. roughly 38%.



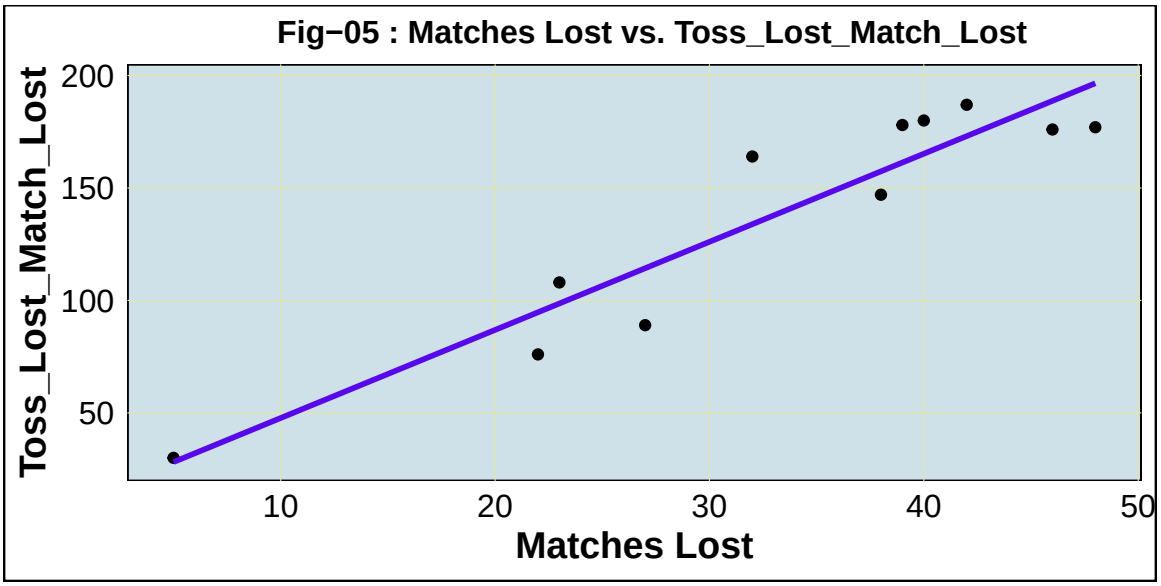
Above plot denotes the "stacked bar graph" of the teams in IPL based on their winning and losing the games. By looking carefully on the plot we find that "Mumbai Indians" have played the maximum no. of matches and has won most matches (represented by gray color).This verifies the teams legacy of winning titles.
Also, from here it is clearly evident that "Gujarat Lions(GL)" have played least no. of matches i.e they have appeared for one or two seasons only.
From this graph we can infer one more strong conclusion that "Royal Challengers Bangalore(RCB)" and "Delhi Daredevils(DD)" though have played all the seasons but had faced several defeats owing to which they haven't won IPL trophies so far.



```r
corr_won = cor(x = winner_stats$Toss_Won_Match_Won, y = winner_stats$Total_Played , use = "everything",
               method=c("pearson") )
corr_won  #Correlation
```

```
## [1] 0.8923139
```

From above plot(Fig-04) it is signified that there exists a *good relationships* between "Toss Winning" and "Match winning".Thus, it advisable to the captains to take wise decision after winning the "toss" i.e to bat or to bowl depending upon circumstances.

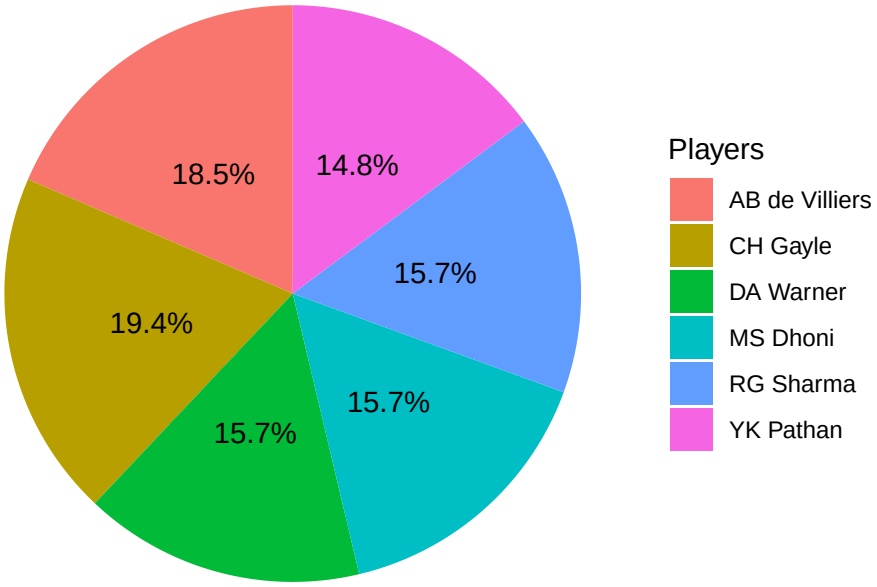



```r
corr_lost_p = cor(x = winner_stats$Toss_Lost_Match_Lost, y = winner_stats$Total_Played, use = "everythi
corr_lost_p
```

```
## [1] 0.939264
```

Now,from above plot(Fig-05) and "correlation" value we cannot say much about losing of the game when a team has lost the toss.

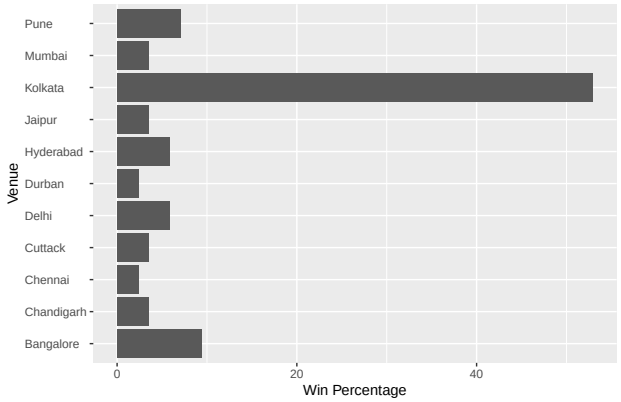
Fig−06 : Pie Chart : MOTM > 15



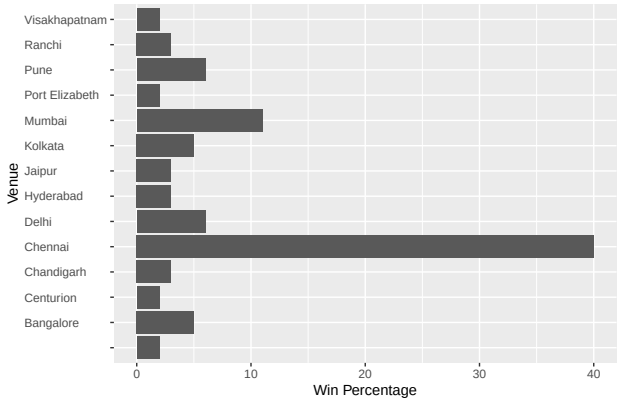
Above unique representation is called as *Pie Diagram*. Here,it represents the percentage of each player about how many times has won 'Man of the Match(MOTM)' title in different IPL matches and has contributed to his team's victory. **Note** : Pie Diagram has been made keeping in mind that 'xyz' player has won 'at least 15' MOTM titles then only he is considered.

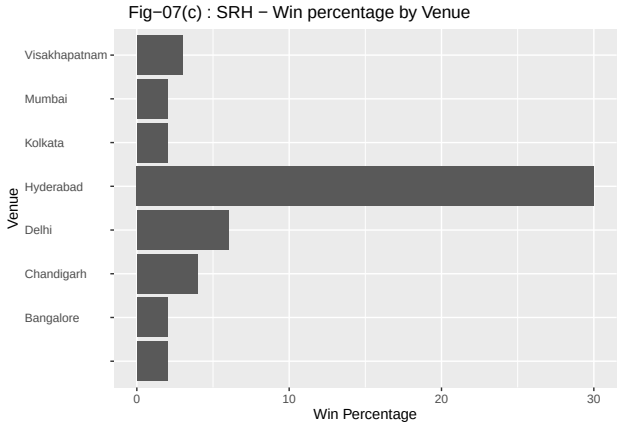
Thus,seeing the pie-chart we can clearly conclude that 'Christopher Henry Gayle' is the best player of IPL since he has won maximum no. of MOTM titles followed by 'Abraham Benjamin De Villiers' who has won many titles and lead to team's victory.

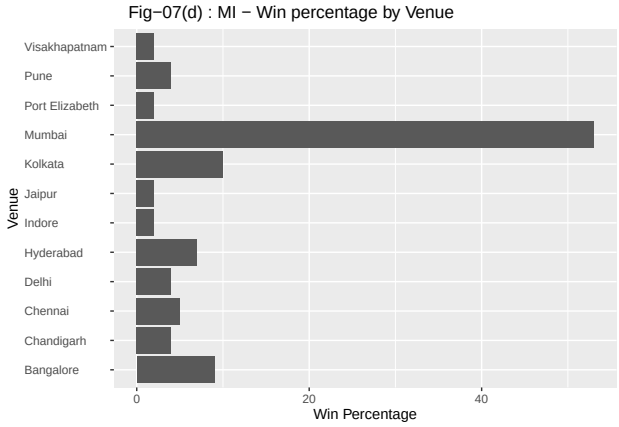
Fig−07(a) : KKR − Win percentage by Venue

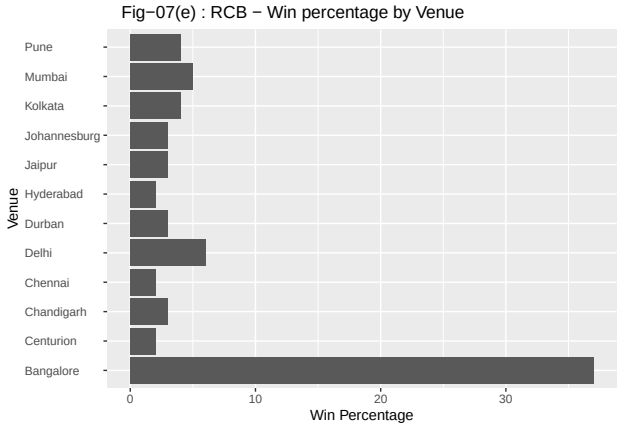


Fig−07(b) : CSK − Win percentage by Venue









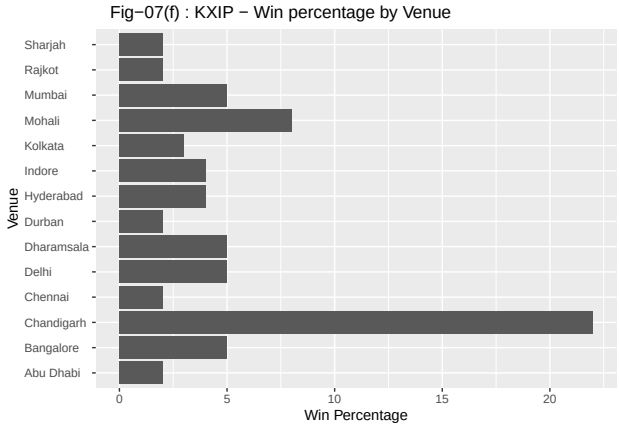
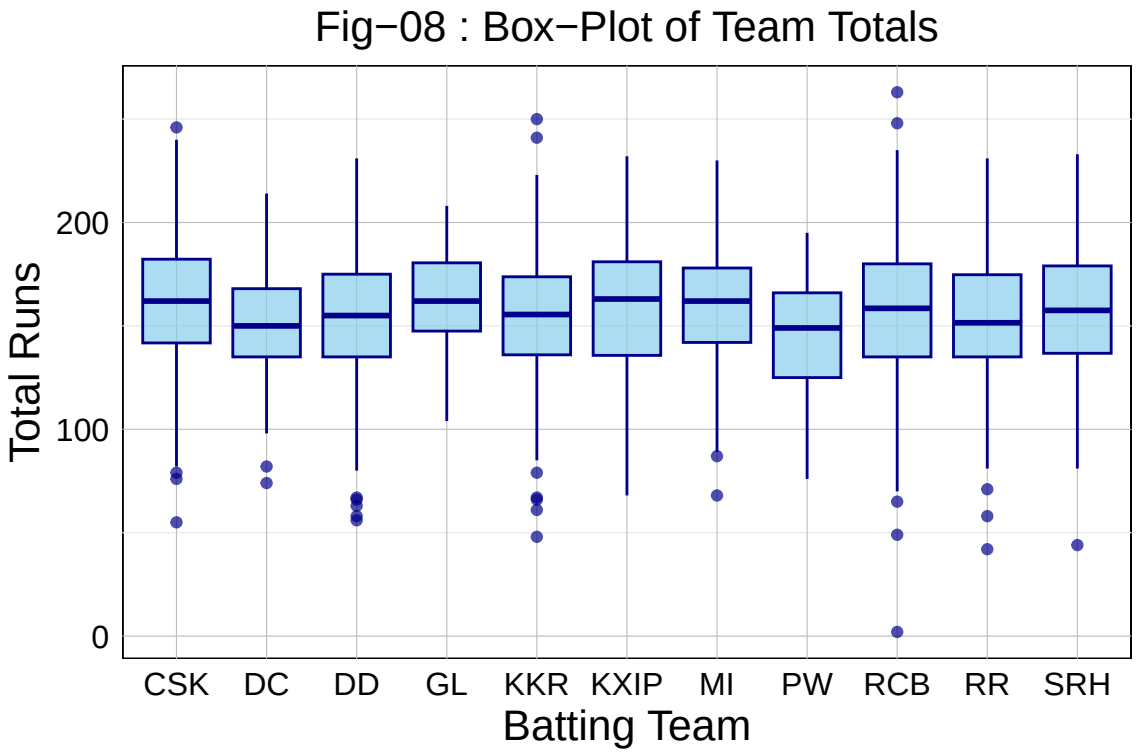
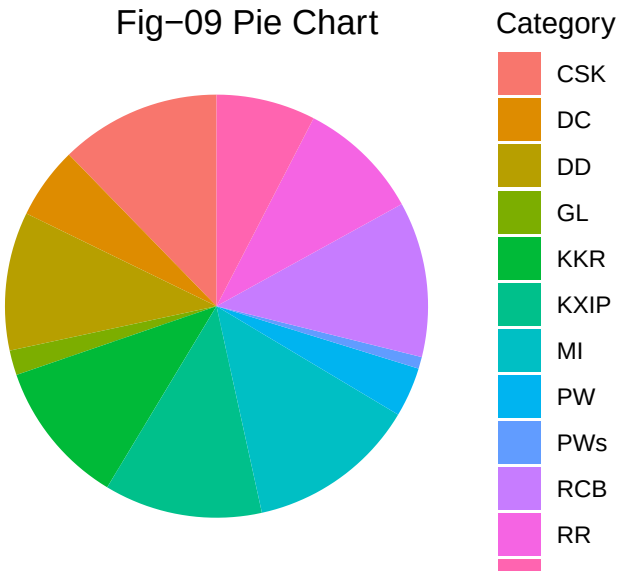


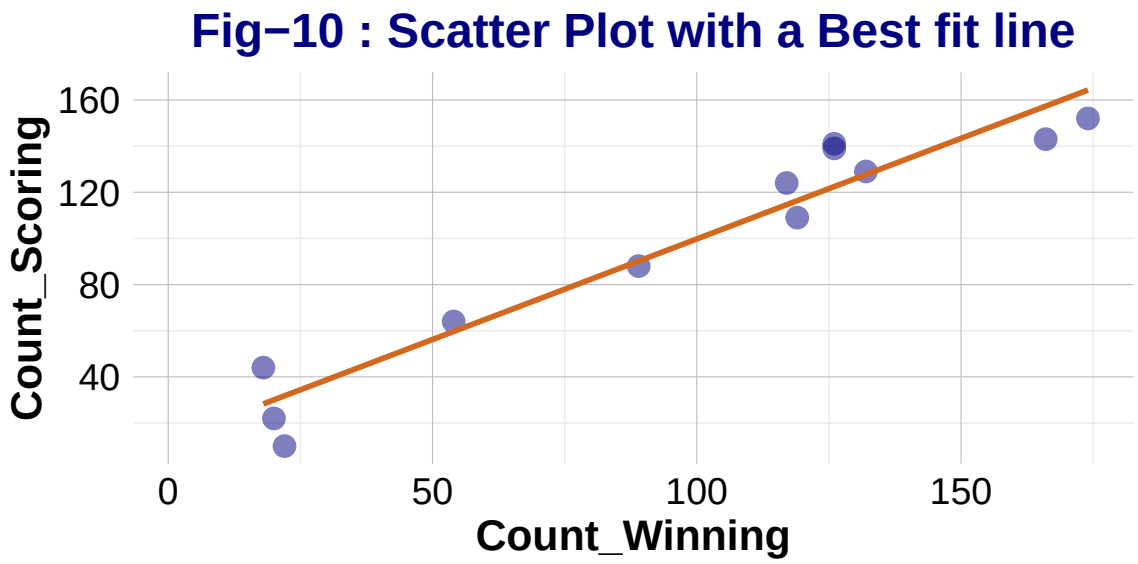
Fig-07(a-f) contains of 6 teams in IPL who have won *(more than 1 matches)* in a particular city and a *horizontal bar graph* of each team has been made.On 'X' axis contains the 'Percentage Win' of a particular team against the cities on 'Y' axis(say KKR in Fig-07(a)). From all of the above prepared graphs it is clearly visible that the teams have won maximum matches at there home grounds such MI has won maximum matches in Mumbai and KXIP has won maximum in Chandigarh,Punjab. Thus,this paves a primary step in *predicting* the teams victory by 'analyzing' the teams previous performance on that venue.



Above Fig-08 represents a 'Box-Plot' which is plotted against the batting teams and their individual totals made in different innings played throughout the IPL.From the plot we can draw some conclusions such as : 1- RCB is the most inconsistent performer as they have more outliers compared to other teams.The team has made 'Maximum Total' and records 'Minimum Total' also in IPL history. 2- KXIP is th most consistent in making the totals i.e they bat almost uniformly in every match and their total ranges in (140-170). similary, many more inferences can be drawn.



Above Fig-09 represents a pie chart of the percentage of how many times teams have made more than thirty runs in Death Overs(16-20). So,we can see that 'CSK' has made 'maximum no. of times' whereas 'SRH' has made 'least number of times'.




```
merged_data = merge(count_winning , count_scoring , by = "Teams" , all = T)
merged_data = na.omit(merged_data)
pearson_corr <- cor(merged_data$Count_Winning, merged_data$Count_Scoring, method = "pearson")
pearson_corr
```

```
## [1] 0.9648518
```

In 'Scatter Plot' *X axis* represents the no. of times the teams have made more than 30 runs whereas *Y axis* represents the no. of times teams have won after *scoring >30 runs*. Thus, from Fig-10 and 'correlation value i.e. 0.96' is clearly visible that the teams which have made more than 30 runs in *deaths overs(16-20)* have won most of the times.

Results

We have come across several results which were obtained using the different rows and colum values of the csv files.

a) **Trends and Patterns**: 1-IPL matches are very high scoring matches with many boundary hits.
2-Teams making death over runs tend to win the match

b) **Key Players and Teams**: 1-Some of the key players observed were CH Gayle, AB De Villiers, David Warner etc.
2- Top 3 Teams - i) - Mumbai Indians(MI)
ii) - Chennai Super Kings(CSK)
iii) - Kolkata Knight Riders(KKR)

c) **Correlations and Relationships**: 1-We have found the +ve correlation between toss winning and match winning.
2-We have found the strong +ve correlation between team securing more than 30 runs in last 4 overs and winning the match.

Conclusion

After playing with the exciting figures of Indian Premier League(2008-19) and visualizing it to greater extent with help of R, we have come across the end of Project. Thus, I would like to conclude as-

a) **Recommendations** : 1-More IPL venues should be introduced in the country as IPL is followed religiously by our countrymen.
2-Mixed gender IPL can be introduced to engage more and more audience.

b) **Limitations** : Some of the data values were missing.If numeric then replaced by median of the data otherwise values were omitted

c) **Future Analytics and Findings** : 1-We can find "World Eleven" i.e the best 11 players.
2- We can make the "Heat-map" of the individual players and its "best zone" and "danger zone" of batting.
3- We can predict the teams total in upcoming according to there previous performance

Thus, in this manner many more exciting analytics can be drawn.