R Notebook

Code

GRIP : The Sparks Foundation

Data Science and Business Analytics

Prepared by : Rutvi Shah

TASK-5 : Exploratory Data Analysis-Sports

Hide

```
library(tidyverse)
library(plyr)
library(dplyr)
```

Hide

```
data_deliveries = read.csv('deliveries.csv',stringsAsFactors = FALSE)
View(data_deliveries)
head(data_deliveries)
```

	match_id <int>	inning <int>	batting_team <chr>	bowling_team <chr>	o… <int>	ball <int>	batsma <chr>
1	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	1	1	DA Warı
2	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	1	2	DA Warı
3	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	1	3	DA Warı
4	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	1	4	DA Warı
5	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	1	5	DA Warı
6	1	1	Sunrisers Hyderabad	Royal Challengers Bangalore	1	6	S Dhaw

6 rows | 1-9 of 21 columns

Hide

```
data_matches = read.csv('matches.csv',stringsAsFactors = FALSE)
View(data_matches)
head(data_matches)
```

	id <int>	Season <chr>	city <chr>	date <chr>	team1 <chr>	team2 <chr>
1	1	IPL-2017	Hyderabad	05-04-2017	Sunrisers Hyderabad	Royal Challengers Bang
2	2	IPL-2017	Pune	06-04-2017	Mumbai Indians	Rising Pune Supergiant
3	3	IPL-2017	Rajkot	07-04-2017	Gujarat Lions	Kolkata Knight Riders
4	4	IPL-2017	Indore	08-04-2017	Rising Pune Supergiant	Kings XI Punjab

	id	Season	city	date	team1	team2
	<int>	<chr>	<chr>	<chr>	<chr>	<chr>
5	5	IPL-2017	Bangalore	08-04-2017	Royal Challengers Bangalore	Delhi Daredevils
6	6	IPL-2017	Hyderabad	09-04-2017	Gujarat Lions	Sunrisers Hyderabad

6 rows | 1-7 of 18 columns

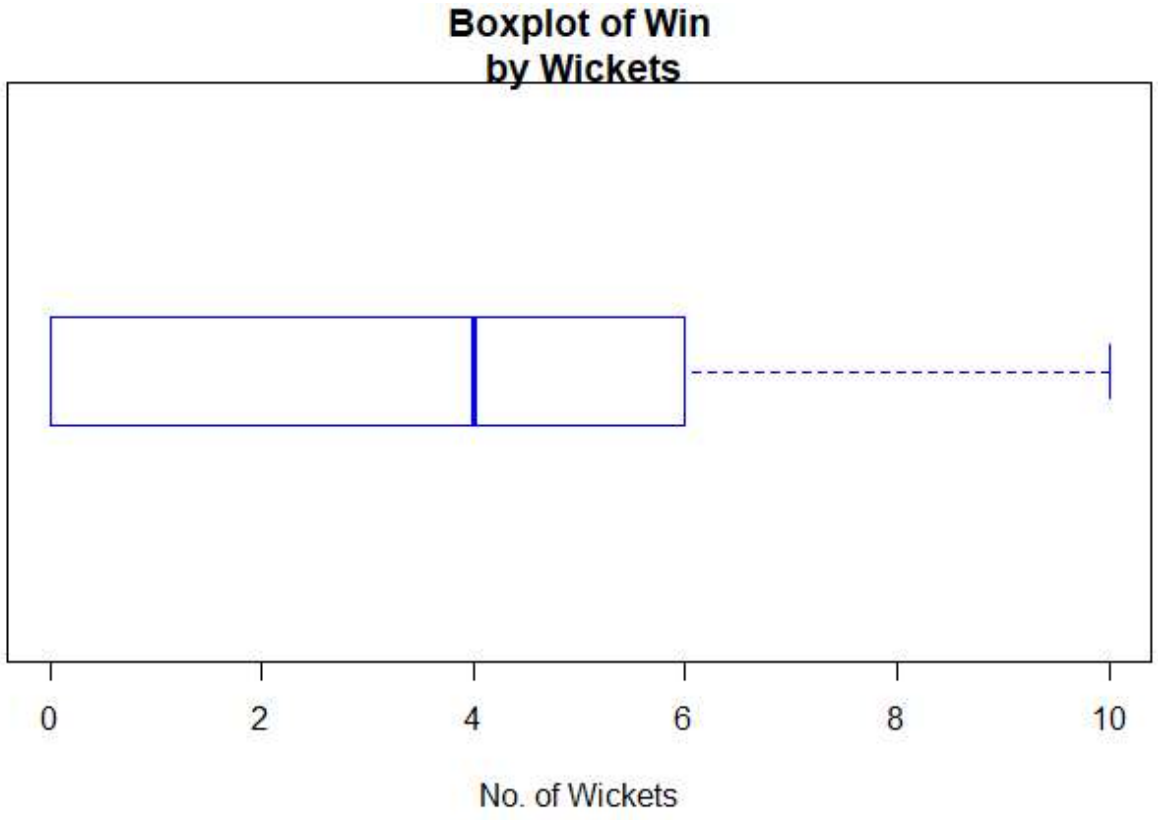
Hide

```
data_matches$Season = as.factor(data_matches$Season)
data_matches$city = as.factor(data_matches$city)
data_matches$team1 = as.factor(data_matches$team1)
data_matches$team2 = as.factor(data_matches$team2)
data_matches$toss_winner = as.factor(data_matches$toss_winner)
data_matches$toss_decision = as.factor(data_matches$toss_decision)
data_matches$result = as.factor(data_matches$result)
data_matches$dl_applied = as.factor(data_matches$dl_applied)
data_matches$winner = as.factor(data_matches$winner)
data_matches$venue = as.factor(data_matches$venue)
data_matches$win_by_runs= as.factor(data_matches$win_by_runs)
```

discriptive statistics

Hide

```
#Measures of Central Tendency
#Boxplot representing the Win by Wickets
boxplot(data_matches$win_by_wickets, horizontal=T, varwidth=TRUE,outline=TRUE, boxwex=0.2, bo
rder=c("blue"),xlab = "No. of Wickets",main="Boxplot of Win
 by Wickets")
```



Hide

```
paste("Mean: ",round(mean(data_matches$win_by_wickets)))
```

```
[1] "Mean:  3"
```

Hide

```
paste("Variance: ",round(var(data_matches$win_by_wickets)))
```

```
[1] "Variance:  11"
```

Hide

```
paste("Standard deviation: ",round(sd(data_matches$win_by_wickets)))
```

```
[1] "Standard deviation:  3"
```

Hide

```
paste("1st Quartile: ",quantile(data_matches$win_by_wickets,prob=c(0.25)))
```

```
[1] "1st Quartile:  0"
```

Hide

```
paste("Median: ",quantile(data_matches$win_by_wickets,prob=c(0.50)))
```

```
[1] "Median:  4"
```

Hide

```
paste("3rd Quartile: ",quantile(data_matches$win_by_wickets,prob=c(0.75)))
```

```
[1] "3rd Quartile:  6"
```

Hide

```
total_season <- length(unique(data_matches$Season))
total_season
```

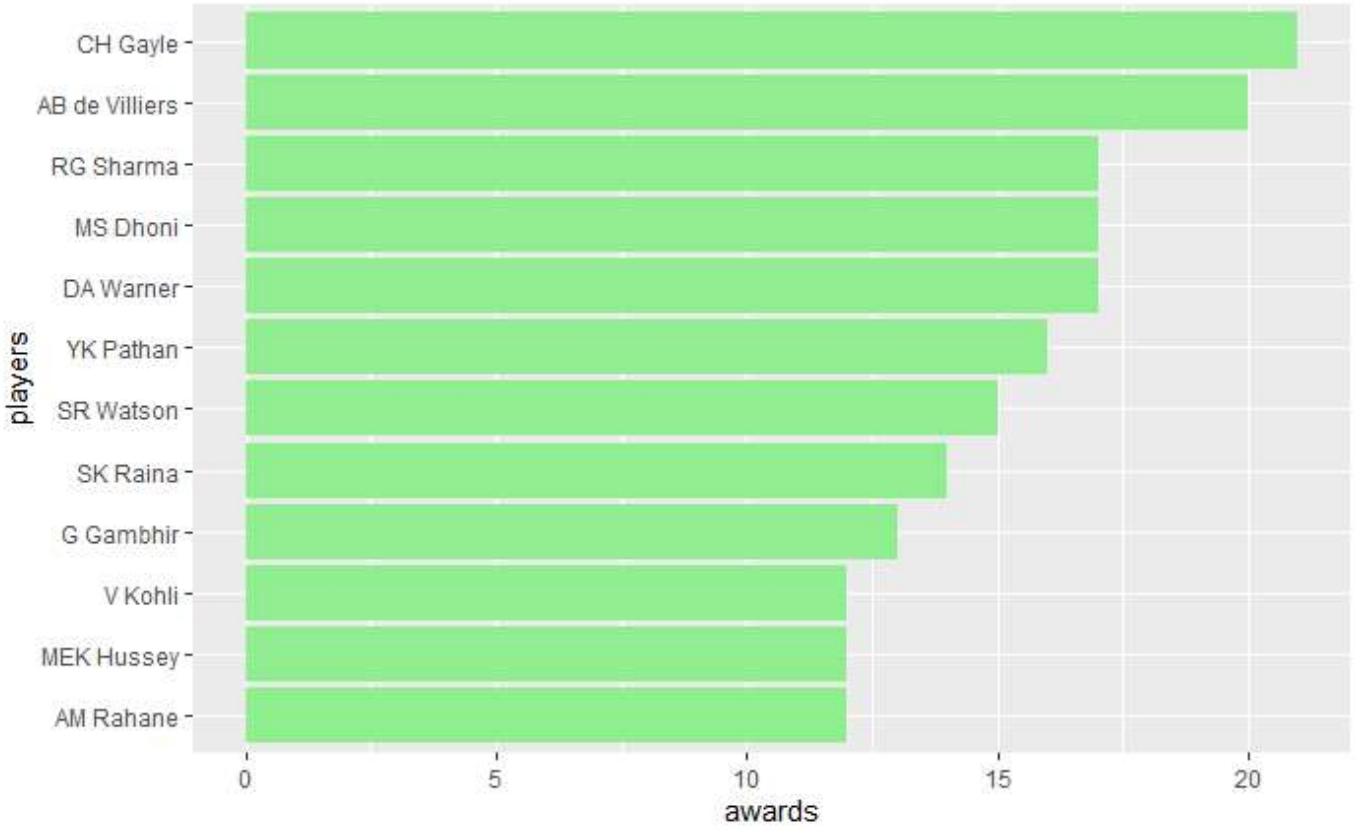
```
[1] 12
```

most has got man of the match

Hide

```
data_matches%>%
  group_by(player_of_match)%>%
  dplyr::summarise(awards =n())%>%
  top_n(10)%>%
  ggplot(aes(x = reorder(player_of_match, awards), y = awards))+
  geom_bar(stat = "identity", fill= "light green")+
  coord_flip()+
  xlab("players")
```

```
Selecting by awards
```



season with most number of matches

Hide

```
data_matches %>%
  group_by(Season)%>%

 dplyr:: summarise(total_match=n())%>%
  filter(total_match==max(total_match))
```

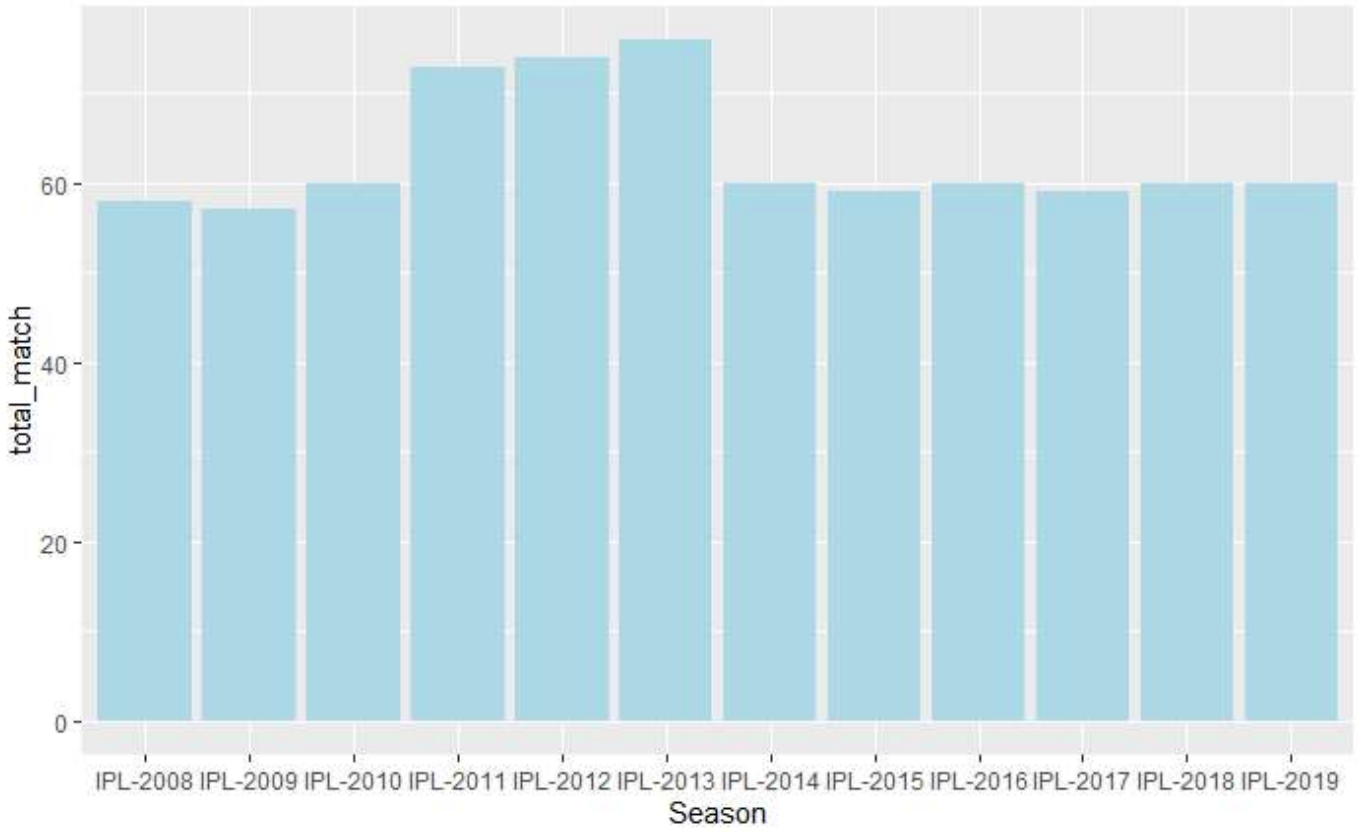
Season	total_match
<fctr>	<int>
IPL-2013	76

1 row

plot graph of season with most number of matches

Hide

```
data_matches%>%
  group_by(Season)%>%
  dplyr::summarise(total_match =n())%>%
ggplot(aes(Season, total_match, fill=Season))+
  geom_bar(stat = "identity", fill="light blue")
```

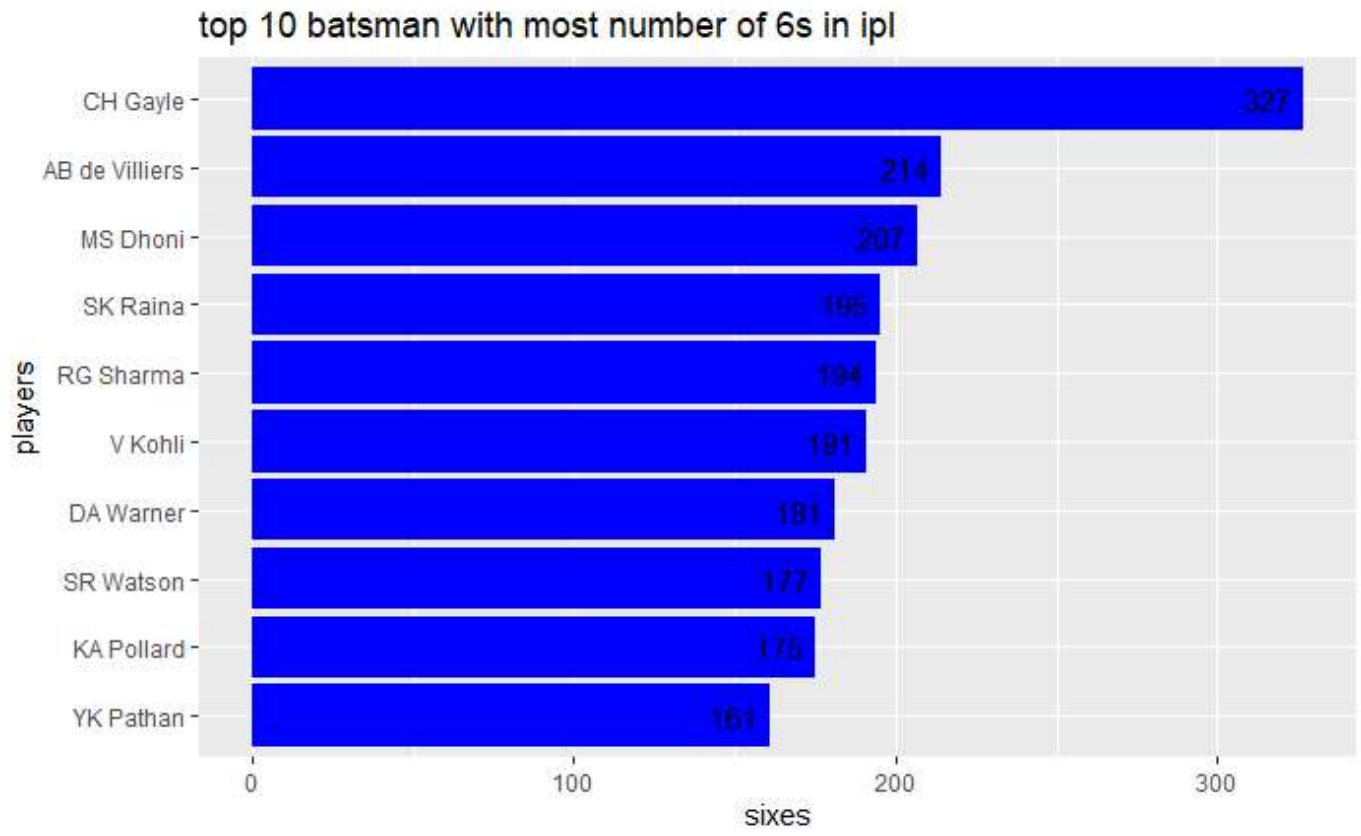


most 6s

Hide

```
data_deliveries %>%
  group_by(batsman) %>%
  filter(batsman_runs == 6) %>%
  dplyr:: summarize(sixes = n()) %>%
  top_n(10) %>%
  ggplot(aes(x = reorder(batsman, sixes), y = sixes))+
  geom_bar(stat = "identity", fill= "blue")+
  coord_flip()+
  xlab("players")+
  ggtitle("top 10 batsman with most number of 6s in ipl")+
geom_text(aes(label = sixes), hjust = 1.25)
```

```
Selecting by sixes
```



which team had won by maximum runs?

Hide

```
maxruns = data_matches[which.max(data_matches$win_by_runs),]
maxruns %>% select('winner', 'win_by_runs','Season')
```

	winner <fctr>	win_by_runs <fctr>	Season <fctr>
44	Mumbai Indians	146	IPL-2017

1 row

which team had won by miximum wicket?

Hide

```
data_matches%>% filter(win_by_wickets==max(win_by_wickets)) %>% select("winner","win_by_wicke
ts","Season")
```

winner <fctr>	win_by_wickets <int>	Season <fctr>
Kolkata Knight Riders	10	IPL-2017
Kings XI Punjab	10	IPL-2017
Deccan Chargers	10	IPL-2008
Delhi Daredevils	10	IPL-2009
Royal Challengers Bangalore	10	IPL-2010
Rajasthan Royals	10	IPL-2011

winner <fctr>	win_by_wickets <int>	Season <fctr>
Mumbai Indians	10	IPL-2012
Chennai Super Kings	10	IPL-2013
Royal Challengers Bangalore	10	IPL-2015
Sunrisers Hyderabad	10	IPL-2016

1-10 of 11 rows Previous 1 2 Next

which team won by minimum wickets?

Hide

```
data_matches %>%
  filter(win_by_wickets != 0) %>%
  filter(win_by_wickets == min(win_by_wickets)) %>%
  select("winner","win_by_wickets","Season")
```

winner <fctr>	win_by_wickets <int>	Season <fctr>
Kolkata Knight Riders	1	IPL-2015
Chennai Super Kings	1	IPL-2018
Sunrisers Hyderabad	1	IPL-2018

3 rows

IMPACT OF TOSS WINNING ON A MATCH

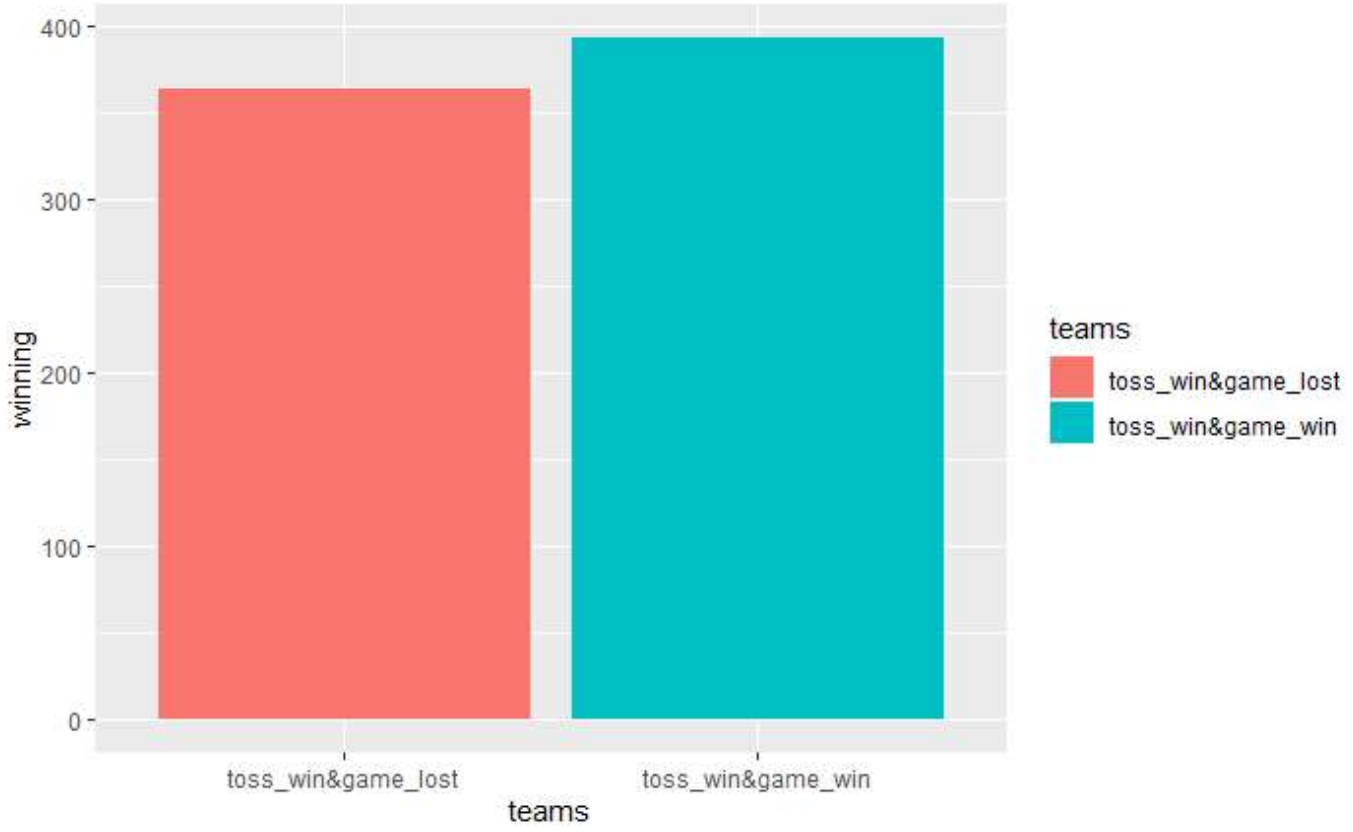
Hide

```
y=0
n=0



for(i in seq(1, nrow(data_matches)))
{
  if(data_matches$toss_winner[i] == data_matches$winner[i])
    y=y+1
  else
    n=n+1
}
if(y>= n)
{
  print(paste("Yes,Toss-winning has an impact of winning a game"))
  print(paste("Matches won by toss winners are:",y, "& Total matches:", nrow(data_matches)))
}
```

```
[1] "Yes,Toss-winning has an impact of winning a game"
[1] "Matches won by toss winners are: 393 & Total matches: 756"
```

Hide

```
winning = c(y,n)
teams = c("toss_win&game_win", "toss_win&game_lost")
df = data.frame(teams,winning,stringsAsFactors = FALSE)

ggplot(df)+geom_bar(aes(teams,winning,fill=teams),stat = "identity")
```



hypothesis 2 : suppose if we opt batting first is help to win the match

Hide

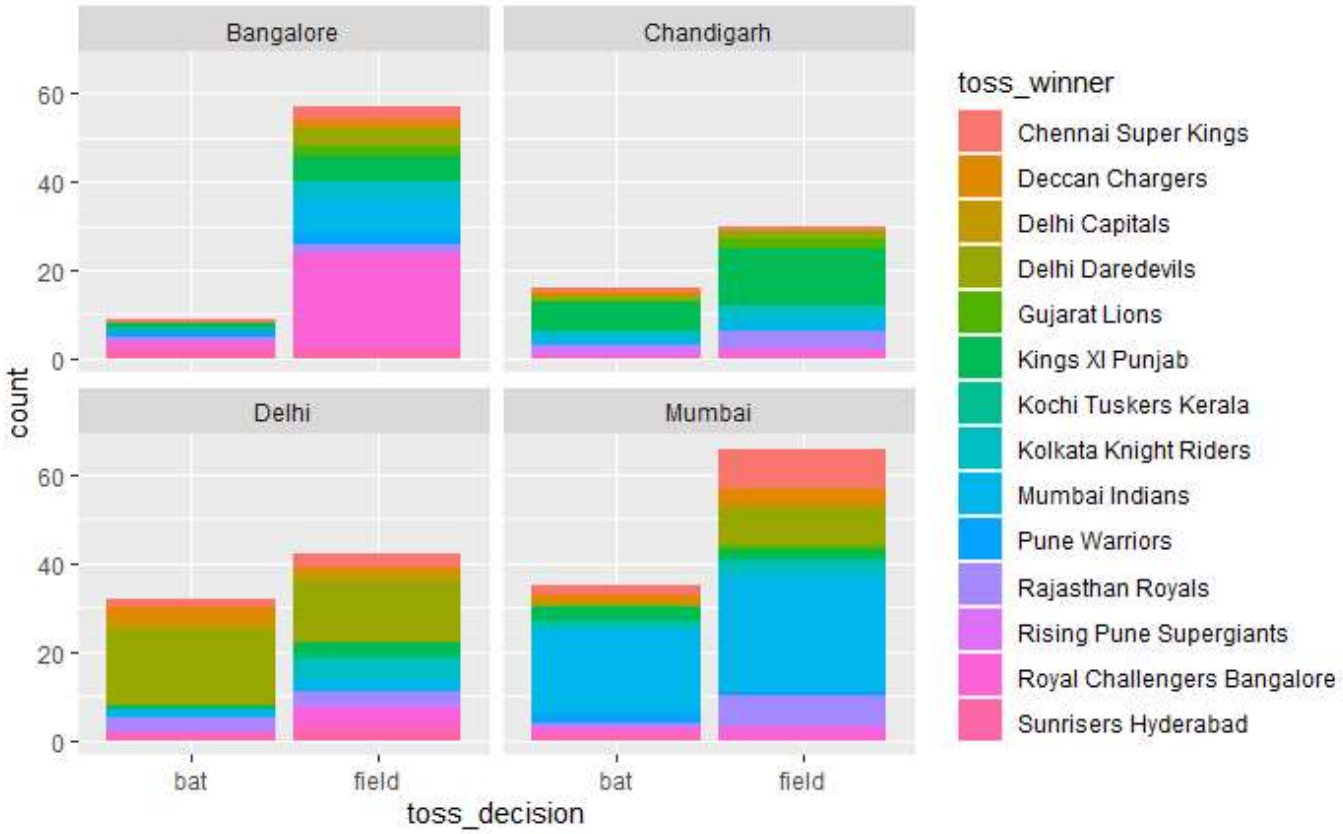
```
toss=data_matches[data_matches$toss_decision=="bat",]
bf=0
bl=0
i=0
for(i in seq(1,nrow(toss)))
{
  if(as.factor(toss$toss_winner[i]==as.character(toss$winner[i])))
  {
    bf=bf+1
  }
  else
  {
    bl=bl+1
  }
}

toss1=data_matches[data_matches$toss_decision=="field",]
fw=0
fl=0
j=0
for(j in seq(1,nrow(toss1)))
{
  if(as.factor(toss1$toss_winner[j]==as.character(toss1$winner[j])))
  {
    fw=fw+1
  }
  else
  {
    fl=fl+1
  }
}
toss_decision = data.frame("Bat first or second"=c("Batting first","fielding first"),"count"=
c(bf,fw))
tibble(toss_decision)
```

Bat.first.or.second <fctr>	count <dbl>
Batting first	293
fielding first	463

2 rows

Hide

```
data_matches%>%filter(city=="Mumbai"|city=="Bangalore"|city=="Delhi"|city=="kolkata"|city=="j
aipur"|city=="Hydrabad"|city=="Chandigarh"|city=="pune")%>% ggplot()+
  geom_bar(aes(x=toss_decision,fill=toss_winner))+facet_wrap(~city)
```



ipl winners

Hide

```
data_matches %>%
  select(Season, id, winner)%>%
  group_by(Season)%>%
  slice(which.max(id))%>%
  select(Season,winner)
```

Season <fctr>	winner <fctr>
IPL-2008	Rajasthan Royals
IPL-2009	Deccan Chargers
IPL-2010	Chennai Super Kings
IPL-2011	Chennai Super Kings
IPL-2012	Kolkata Knight Riders
IPL-2013	Mumbai Indians
IPL-2014	Kolkata Knight Riders
IPL-2015	Mumbai Indians
IPL-2016	Sunrisers Hyderabad
IPL-2017	Mumbai Indians

1-10 of 12 rows Previous **1** 2 Next

Highest total individual

Hide

```
data_deliveries%>%
  group_by(bowling_team, batsman)%>%
 dplyr:: summarise(total_runs=sum(batsman_runs))%>%
  arrange(total_runs)%>%
  na.omit()%>%
  group_by(bowling_team)%>%
  slice(which.max(total_runs))
```

```
`summarise()` has grouped output by 'bowling_team'. You can override using the `.groups` argu
ment.
```

bowling_team <chr>	batsman <chr>	total_runs <int>
Chennai Super Kings	V Kohli	749
Deccan Chargers	R Dravid	339
Delhi Capitals	AD Russell	118
Delhi Daredevils	V Kohli	763
Gujarat Lions	DA Warner	336
Kings XI Punjab	DA Warner	833
Kochi Tuskers Kerala	SR Tendulkar	100
Kolkata Knight Riders	DA Warner	835
Mumbai Indians	SK Raina	824
Pune Warriors	CH Gayle	383

1-10 of 15 rows Previous 1 2 Next

Hide

```
data_deliveries %>%
  group_by(dismissal_kind,fielder) %>%
  dplyr::summarise(total= sum(table(dismissal_kind))) %>%
  arrange(total) %>%
  group_by(dismissal_kind) %>%
  slice(which.max(total)) %>% na.omit()
```

```
`summarise()` has grouped output by 'dismissal_kind'. You can override using the `.groups` ar
gument.
```

dismissal_kind <chr>	fielder <chr>	total <int>
		170244
bowled		1581
caught	KD Karthik	109

dismissal_kind <chr>	fielder <chr>	total <int>
caught and bowled		211
hit wicket		10
lbw		540
obstructing the field		2
retired hurt		12
run out	MS Dhoni	23
stumped	MS Dhoni	38

1-10 of 10 rows

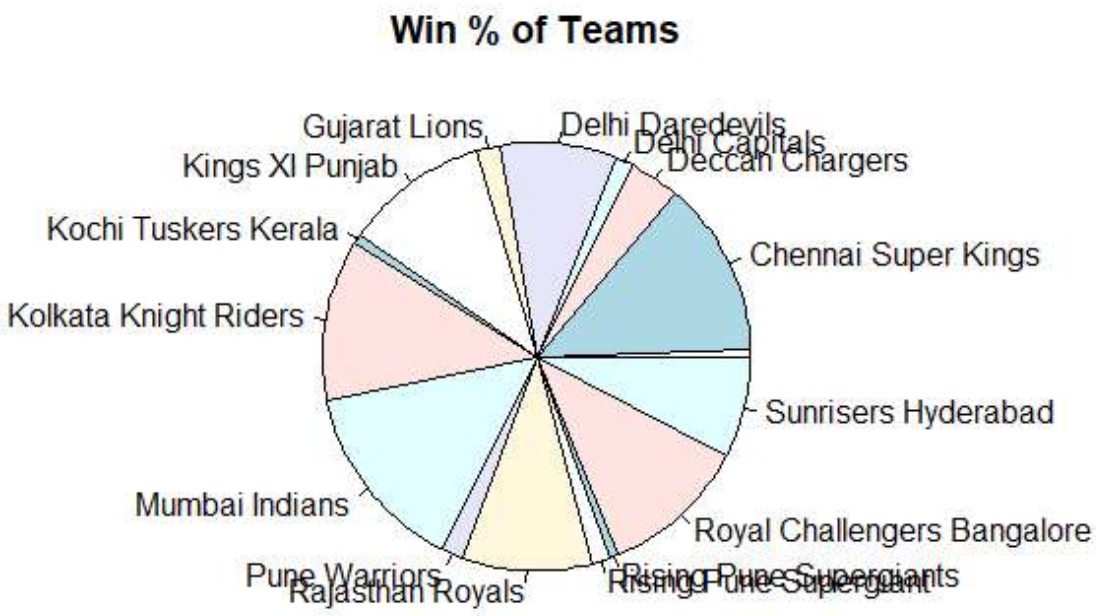
Hide

```
library(corrplot)
```

```
package 'corrplot' was built under R version 3.6.3corrplot 0.84 loaded
```

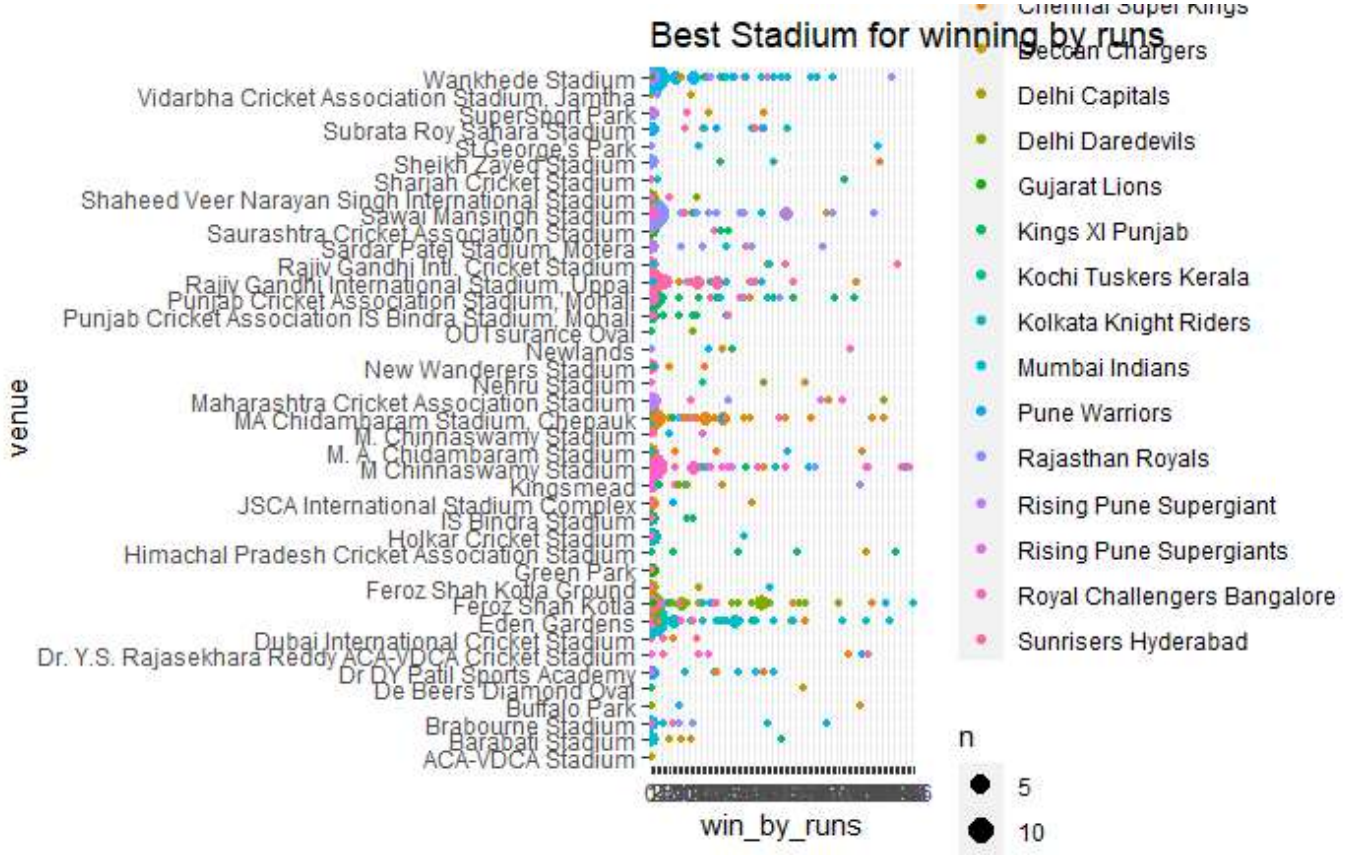
Hide

```
x = data.frame(data_matches$Season,data_matches$win_by_runs,data_matches$win_by_wickets,data_matches$dl_applied)
pie(table(data_matches$winner), main="Win % of Teams")
```



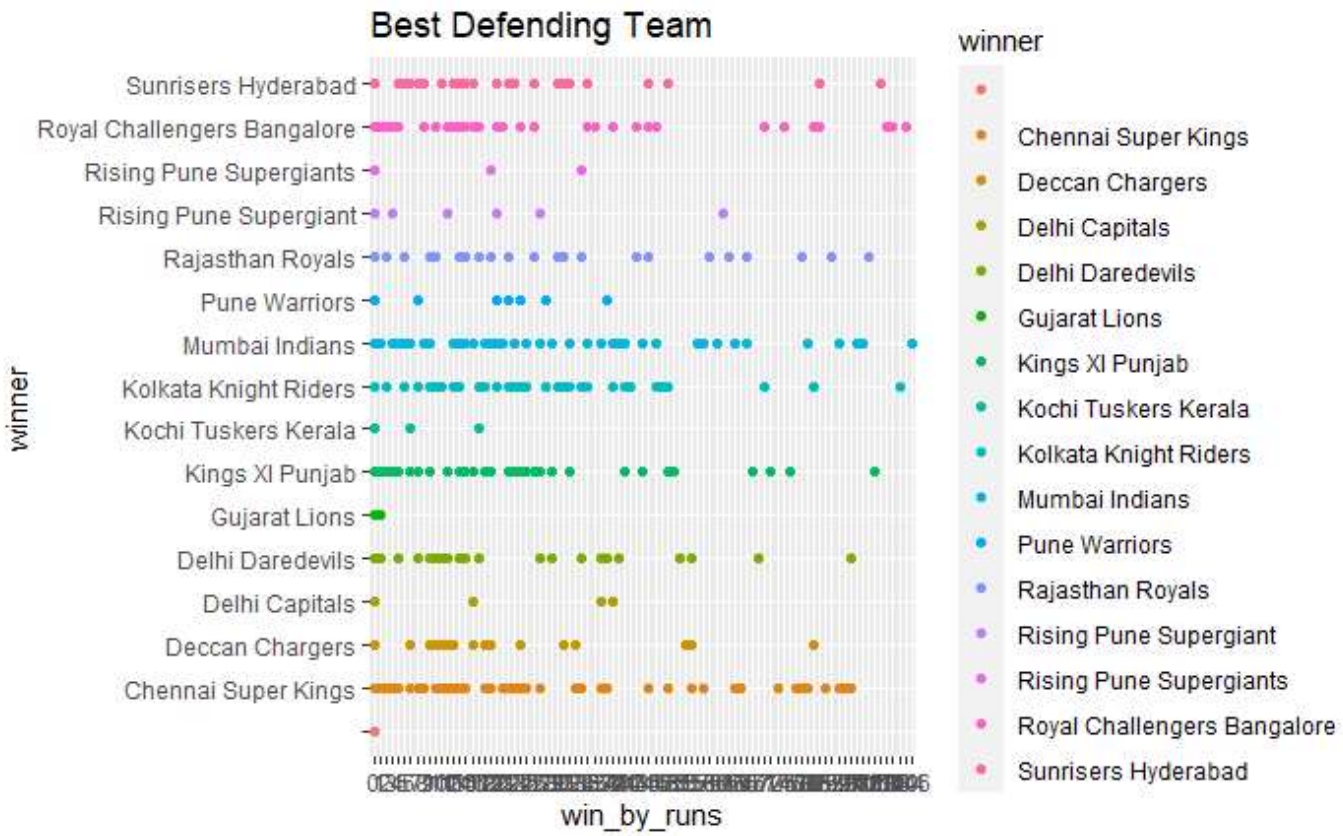
Hide

```
#Scatter Plot representing the Best stadiums to win by a large margin of runs
library(ggplot2)
ggplot(data_matches,aes(win_by_runs, venue, colour = winner)) + geom_count() +
ggtitle("Best Stadium for winning by runs")
```



Hide

```
ggplot(data_matches,aes(win_by_runs, winner, colour = winner)) + geom_point() +
ggtitle("Best Defending Team")
```


Best Defending Team
winner
Sunrisers Hyderabad
Royal Challengers Bangalore
Rising Pune Supergiants
Rising Pune Supergiant
Rajasthan Royals
Pune Warriors
Mumbai Indians
Kolkata Knight Riders
Kochi Tuskers Kerala
Kings XI Punjab
Gujarat Lions
Delhi Daredevils
Delhi Capitals
Deccan Chargers
Chennai Super Kings
win_by_runs
winner
Chennai Super Kings
Deccan Chargers
Delhi Capitals
Delhi Daredevils
Gujarat Lions
Kings XI Punjab
Kochi Tuskers Kerala
Kolkata Knight Riders
Mumbai Indians
Pune Warriors
Rajasthan Royals
Rising Pune Supergiant
Rising Pune Supergiants
Royal Challengers Bangalore
Sunrisers Hyderabad