Project-1, ST558

Pratap Adhikari

9/18/2020

Contents

Project1-ST558

This project work involves creating a `vignette` for reading and summarizing data from the *National Hockey League's* (NHL) **API**.

List of library packages

The list of library packages I have used to run this code in order to carry on this project are:

- knitr
- httr
- jsonlite
- tidyverse
- dplyr
- haven
- ggplot2
- qwraps2
- rmarkdown
- RSQlite

To install the packages: install.packages("knitr", "httr", "jsonlite", "tidyverse", "dplyr", "haven", "ggplot2", "qwraps2", "rmarkdown", "RSQlite")

Function to get franchiseAPI

```
#create funciton to read data from records API
nhl<- function(tabName, ID=NULL, ...){
  base_url<- "https://records.nhl.com/site/api"
  if (!is.null(tabName)){

     if ( tabName %in% c("franchise", "franchise-team-totals") && (!is.null(ID))){
    stop("This tab can not return with 'ID' defined")
  }

    if (is.null(ID)){
    full_url<- paste0(base_url, "/", tabName)
  }

    if (!is.null(ID)){
     full_url<- paste0(base_url, "/", tabName, ID)
  }
  get_nfl<- GET(full_url)
```

```
  txt_nfl<- content(get_nfl, "text")  # convert to JSON text form
  json_nfl<- fromJSON(txt_nfl, flatten=T) # convert to list
return(json_nfl)
  }
  else {
    return("Invalid tabName")
  }
}
```

Function to get statsAPI

```
#create funciton to data from statsAPI
nhl_modifier<- function(modifier, ID=NULL,...){
  stbase_url<- "https://statsapi.web.nhl.com/api/v1/teams"
  if (modifier %in%  c("expand=team.roster", "expand=person.names", "expand=team.schedule.next", "expand
  {
    get_st<- GET(paste0(stbase_url, "?", modifier))
st_txt<- content(get_st, "text")
json_st<- fromJSON(st_txt, flatten=T)
}
else { #return a message if the modifier is not campatible with the function
  json_st="Sorry, can't accept this modifier"

}
return(json_st)
}
```

```
nhlData<- function (tabName=NULL, modifier=NULL, ID=NULL, ...){
if (!is.null(tabName) && !is.null(modifier)){
  stop("it can not work together with tabName and modifier")

}
if(is.null(tabName) && is.null(modifier) ){

  output<- nhl("franchise")
}
  if (is.null(modifier) && !is.null(tabName)){
     output<- nhl(tabName, ID)
     output<- output$data
  }
  # if modifier is not null and id is null
  if(!is.null(modifier) && is.null(id)){
   output<- nhl_modifier(modifier)
  }

  #if both modifier and id are not  null
  if(!is.null(modifier) && !is.null(id) ){
    output<- nhl_modifier(modifier)
    output<- output$teams
    output<- output %>% filter(id==ID) %>% select(id:roster.link)
```

```
  }
  return(output)
}
```

Overview of franchise and location

```
teamtotal<- nhl(tabName = "franchise-team-totals")$data
```

```
## No encoding supplied: defaulting to UTF-8.
```

```
#getteams from another endpoint
division<- nhl_modifier(modifier = "expand=team.roster")$teams %>% select(id, division.name,  locationNa
#join the two dataset from two different APIs
newData<- left_join(teamtotal, division, by="teamId")
head(newData, n=4)
```

```
##   id activeFranchise firstSeasonId franchiseId gameTypeId gamesPlayed
## 1  1               1      19821983          23          2        2937
## 2  2               1      19821983          23          3         257
## 3  3               1      19721973          22          2        3732
## 4  4               1      19721973          22          3         294
##   goalsAgainst goalsFor homeLosses homeOvertimeLosses homeTies homeWins
## 1         8708     8647        507                 82       96      783
## 2          634      697         53                  0       NA       74
## 3        11779    11889        674                 81      170      942
## 4          857      935         50                  1       NA       90
##   lastSeasonId losses overtimeLosses penaltyMinutes pointPctg points roadLosses
## 1           NA   1181            162          44397    0.5330   3131        674
## 2           NA    120              0           4266    0.0039      2         67
## 3           NA   1570            159          57422    0.5115   3818        896
## 4           NA    133              0           5564    0.0136      8         83
##   roadOvertimeLosses roadTies roadWins shootoutLosses shootoutWins shutouts
## 1                 80      123      592             79           78      193
## 2                  0       NA       63              0            0       25
## 3                 78      177      714             67           82      167
## 4                  2       NA       71              0            0       12
##   teamId           teamName ties triCode wins division.name locationName
## 1      1  New Jersey Devils  219     NJD 1375  Metropolitan   New Jersey
## 2      1  New Jersey Devils   NA     NJD  137  Metropolitan   New Jersey
## 3      2 New York Islanders  347     NYI 1656  Metropolitan     New York
## 4      2 New York Islanders   NA     NYI  161  Metropolitan     New York
##   division.nameShort conference.name
## 1              Metro         Eastern
## 2              Metro         Eastern
## 3              Metro         Eastern
## 4              Metro         Eastern
```

```
# overview of after joining two datasets from two different API endpoints
kable(newData %>% select(id, franchiseId, teamName, locationName) , caption= "Franchise ID, Team Name, L
```

Table 1: Franchise ID, Team Name, Location table for your reference:

id	franchiseId	teamName	locationName
1	23	New Jersey Devils	New Jersey
2	23	New Jersey Devils	New Jersey
3	22	New York Islanders	New York
4	22	New York Islanders	New York
5	10	New York Rangers	New York
6	10	New York Rangers	New York
7	16	Philadelphia Flyers	Philadelphia
8	16	Philadelphia Flyers	Philadelphia
9	17	Pittsburgh Penguins	Pittsburgh
10	17	Pittsburgh Penguins	Pittsburgh
11	6	Boston Bruins	Boston
12	6	Boston Bruins	Boston
13	19	Buffalo Sabres	Buffalo
14	19	Buffalo Sabres	Buffalo
15	1	Montréal Canadiens	Montréal
16	1	Montréal Canadiens	Montréal
17	30	Ottawa Senators	Ottawa
18	30	Ottawa Senators	Ottawa
19	5	Toronto Maple Leafs	Toronto
20	5	Toronto Maple Leafs	Toronto
21	35	Atlanta Thrashers	NA
22	35	Atlanta Thrashers	NA
23	26	Carolina Hurricanes	Carolina
24	26	Carolina Hurricanes	Carolina
25	33	Florida Panthers	Florida
26	33	Florida Panthers	Florida
27	31	Tampa Bay Lightning	Tampa Bay
28	31	Tampa Bay Lightning	Tampa Bay
29	24	Washington Capitals	Washington
30	24	Washington Capitals	Washington
31	11	Chicago Blackhawks	Chicago
32	11	Chicago Blackhawks	Chicago
33	12	Detroit Red Wings	Detroit
34	12	Detroit Red Wings	Detroit
35	34	Nashville Predators	Nashville
36	34	Nashville Predators	Nashville
37	18	St. Louis Blues	St. Louis
38	18	St. Louis Blues	St. Louis
39	21	Calgary Flames	Calgary
40	21	Calgary Flames	Calgary
41	27	Colorado Avalanche	Colorado
42	27	Colorado Avalanche	Colorado
43	25	Edmonton Oilers	Edmonton
44	25	Edmonton Oilers	Edmonton
45	20	Vancouver Canucks	Vancouver
46	20	Vancouver Canucks	Vancouver
47	32	Anaheim Ducks	Anaheim
48	32	Anaheim Ducks	Anaheim
49	15	Dallas Stars	Dallas
50	15	Dallas Stars	Dallas

id	franchiseId	teamName	locationName
51	14	Los Angeles Kings	Los Angeles
52	14	Los Angeles Kings	Los Angeles
53	28	Phoenix Coyotes	NA
54	28	Phoenix Coyotes	NA
55	29	San Jose Sharks	San Jose
56	29	San Jose Sharks	San Jose
57	36	Columbus Blue Jackets	Columbus
58	36	Columbus Blue Jackets	Columbus
59	37	Minnesota Wild	Minnesota
60	37	Minnesota Wild	Minnesota
61	15	Minnesota North Stars	NA
62	15	Minnesota North Stars	NA
63	27	Quebec Nordiques	NA
64	27	Quebec Nordiques	NA
65	28	Winnipeg Jets (1979)	NA
66	28	Winnipeg Jets (1979)	NA
67	26	Hartford Whalers	NA
68	26	Hartford Whalers	NA
69	23	Colorado Rockies	NA
70	23	Colorado Rockies	NA
71	3	Ottawa Senators (1917)	NA
72	3	Ottawa Senators (1917)	NA
73	4	Hamilton Tigers	NA
74	9	Pittsburgh Pirates	NA
75	9	Pittsburgh Pirates	NA
76	9	Philadelphia Quakers	NA
77	12	Detroit Cougars	NA
78	12	Detroit Cougars	NA
79	2	Montreal Wanderers	NA
80	4	Quebec Bulldogs	NA
81	7	Montreal Maroons	NA
82	7	Montreal Maroons	NA
83	8	New York Americans	NA
84	8	New York Americans	NA
85	3	St. Louis Eagles	NA
86	13	Oakland Seals	NA
87	13	Oakland Seals	NA
88	21	Atlanta Flames	NA
89	21	Atlanta Flames	NA
90	23	Kansas City Scouts	NA
91	13	Cleveland Barons	NA
92	12	Detroit Falcons	NA
93	12	Detroit Falcons	NA
94	8	Brooklyn Americans	NA
95	35	Winnipeg Jets	Winnipeg
96	35	Winnipeg Jets	Winnipeg
97	28	Arizona Coyotes	Arizona
98	38	Vegas Golden Knights	Vegas
99	38	Vegas Golden Knights	Vegas
100	13	California Golden Seals	NA
101	5	Toronto Arenas	NA
102	5	Toronto Arenas	NA

id	franchiseId	teamName	locationName
103	5	Toronto St. Patricks	NA
104	5	Toronto St. Patricks	NA
105	28	Arizona Coyotes	Arizona

Read table from two different APIs

```
dta1<- nhlData("franchise")
```

```
## No encoding supplied: defaulting to UTF-8.
```

```
dta1<- dta1 %>% select(id, mostRecentTeamId, teamCommonName, teamPlaceName)
golietable15<- nhl(tabName  = "franchise-goalie-records?cayenneExp=franchiseId=", ID=15)$data %>% selec
```

```
## No encoding supplied: defaulting to UTF-8.
```

Analysis on team ID=15 (Dallas Stars)

```
#Create new variable by adding first and last name from two different columns
golietable15$playerName<- c(paste0(golietable15$firstName, " ", golietable15$lastName))
#select only the varaibles required to analyse the data
golietable15<- golietable15 %>% select(franchiseName, playerName, playerId, activePlayer, gameTypeId, g
```

Categorical Summary Table showing active players from the **Dallas** team

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
 freqtbl<- table (golietable15  %>% group_by(activePlayer)%>% select(franchiseName, activePlayer))
freqtbl
```

```
##               activePlayer
## franchiseName  FALSE TRUE
##   Dallas Stars    34    3
```

```
add_header_above(header = c("Franchise" = 1, "Active Player" = 2),   kable(freqtbl))
```

Franchise	Active Player	
	FALSE	TRUE
Dallas Stars	34	3

```
# Max goals against one game
kable( table(golietable15 %>% group_by(mostGoalsAgainstOneGame) %>% select(mostGoalsAgainstOneGame, pla
```

	Alex Auld	Allan Bester	Anders Lindback	Andrew Raycroft	Andy Moog	Anton Khudobin	Antti Niemi	
3	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	
5	1	1	0	0	0	0	0	
6	0	0	1	0	0	1	1	
7	0	0	0	1	0	0	0	
8	0	0	0	0	0	0	0	
10	0	0	0	0	1	0	0	

```
#create new variables
wlRate<- golietable15 %>% mutate(tiesRate= round(ties/gamesPlayed, 2), winRate= round(wins/gamesPlayed,
kable(wlRate)
```

Win/Loss Rate

playerName	playerId	gameTypeId	gamesPlayed	activePlayer	winRate	lossRate
Don Beaupre	8445381	2	315	FALSE	0.40	0.40
Cesare Maniago	8450020	2	420	FALSE	0.35	0.45
Marty Turco	8460612	2	509	FALSE	0.51	0.30
Kari Lehtonen	8470140	2	445	FALSE	0.49	0.34
Ed Belfour	8445386	2	307	FALSE	0.52	0.31
Allan Bester	8445458	2	10	FALSE	0.40	0.50
Daniel Berthiaume	8445462	2	5	FALSE	0.20	0.60
Gary Edwards	8446602	2	51	FALSE	0.29	0.35
Brian Hayward	8447701	2	26	FALSE	0.23	0.58
Jean Levasseur	8448807	2	1	FALSE	0.00	1.00
Markus Mattsson	8449291	2	2	FALSE	0.50	0.50
Roland Melanson	8449547	2	26	FALSE	0.27	0.42
Gilles Meloche	8449550	2	327	FALSE	0.43	0.36
Lindsay Middlebrook	8449588	2	3	FALSE	0.00	0.00
Andy Moog	8449681	2	175	FALSE	0.43	0.37
Gump Worsley	8450152	2	107	FALSE	0.36	0.35
Gary Smith	8451528	2	39	FALSE	0.26	0.49
Ron Tugnutt	8451837	2	42	FALSE	0.43	0.40
Darcy Wakaluk	8452248	2	65	FALSE	0.35	0.48
Arturs Irbe	8456692	2	35	FALSE	0.49	0.34
Roman Turek	8458266	2	55	FALSE	0.55	0.25
Corey Hirsch	8458680	2	2	FALSE	0.00	0.50
Tim Thomas	8460703	2	8	FALSE	0.25	0.50
Johan Hedberg	8460704	2	19	FALSE	0.63	0.21
Johan Holmqvist	8466303	2	2	FALSE	0.50	0.00
Andrew Raycroft	8467453	2	29	FALSE	0.34	0.45
Alex Auld	8467913	2	21	FALSE	0.43	0.29
Brent Krahn	8468489	2	1	FALSE	0.00	0.00
Mike McKenna	8470093	2	2	FALSE	0.50	0.50
Ben Bishop	8471750	2	143	TRUE	0.52	0.34
Jhonas Enroth	8473523	2	13	FALSE	0.38	0.38
Antti Niemi	8474550	2	85	FALSE	0.44	0.29
Anders Lindback	8474765	2	10	FALSE	0.20	0.80
Jussi Rynnas	8475680	2	2	FALSE	0.00	0.50
Cristopher Nilstorp	8476846	2	6	FALSE	0.17	0.50
Mike Smith	8469608	2	44	TRUE	0.55	0.32
Anton Khudobin	8471418	2	71	TRUE	0.45	0.35

Sumamry table Numeric Summary

```
sumry<- function (x, ...){
  dta<- wlRate %>% filter(gameTypeId == x) %>% select(winRate, lossRate)
  if (x==2) type<- "regular season" else type<- "play off season"
  kable (apply(dta, 2, summary), format="html", digit =4, caption = paste0("Summary among all of the pla
}
# Regular season summary
sumry(2)
```

Summary among all of the players

winRate

lossRate

Min.

0.0000

0.0000

1st Qu.

0.2500

0.3200

Median

0.4000

0.3800

Mean

0.3465

0.3978

3rd Qu.

0.4900

0.5000

Max.

0.6300

1.0000

Sumamry of number of games played

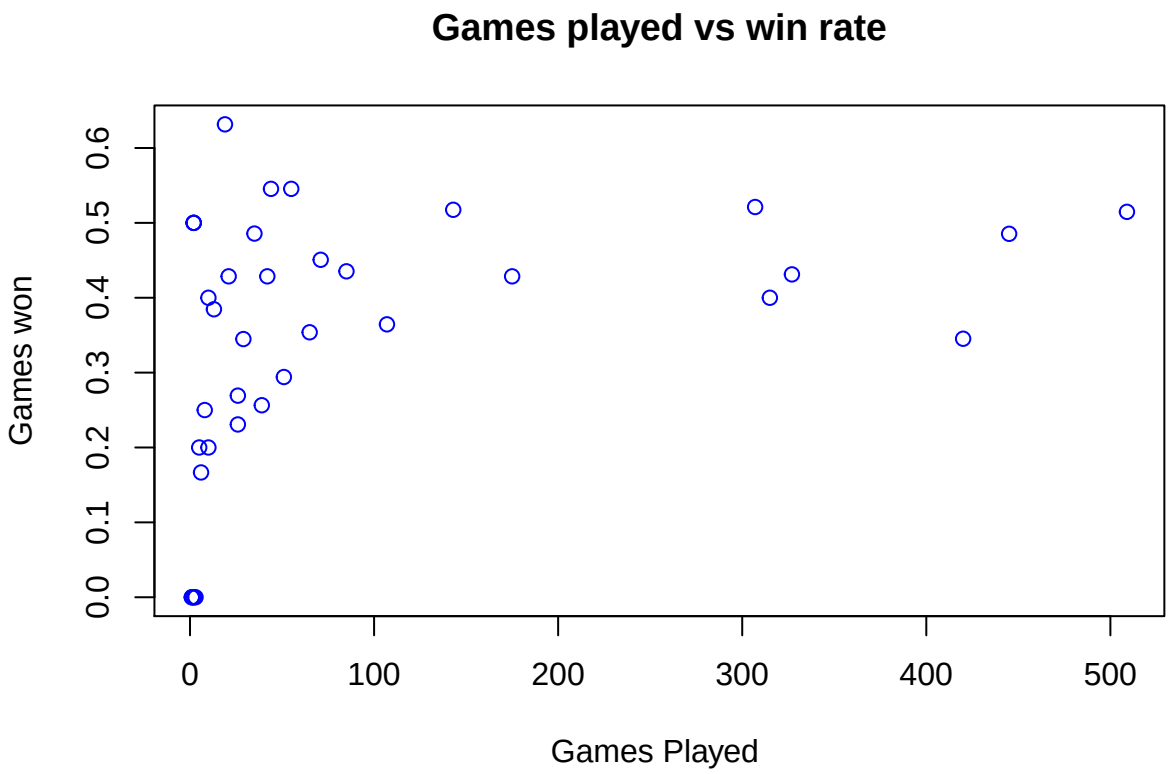
```
gpl<- golietable15 %>% select(gamesPlayed, wins)
kable (apply(gpl, 2, summary) )
```

	gamesPlayed	wins
Min.	1.00000	0.0000
1st Qu.	6.00000	1.0000
Median	29.00000	10.0000
Mean	92.51351	40.7027
3rd Qu.	85.00000	37.0000
Max.	509.00000	262.0000

Plots Every plot I am trying to show the plot difference and advanced options available to plot the same plots. ##### Scatter

The higest win rate is the individuals who have played less games. But importently, higher number of games player looks like they are more consistent in winnig rate than the lower number of games player.

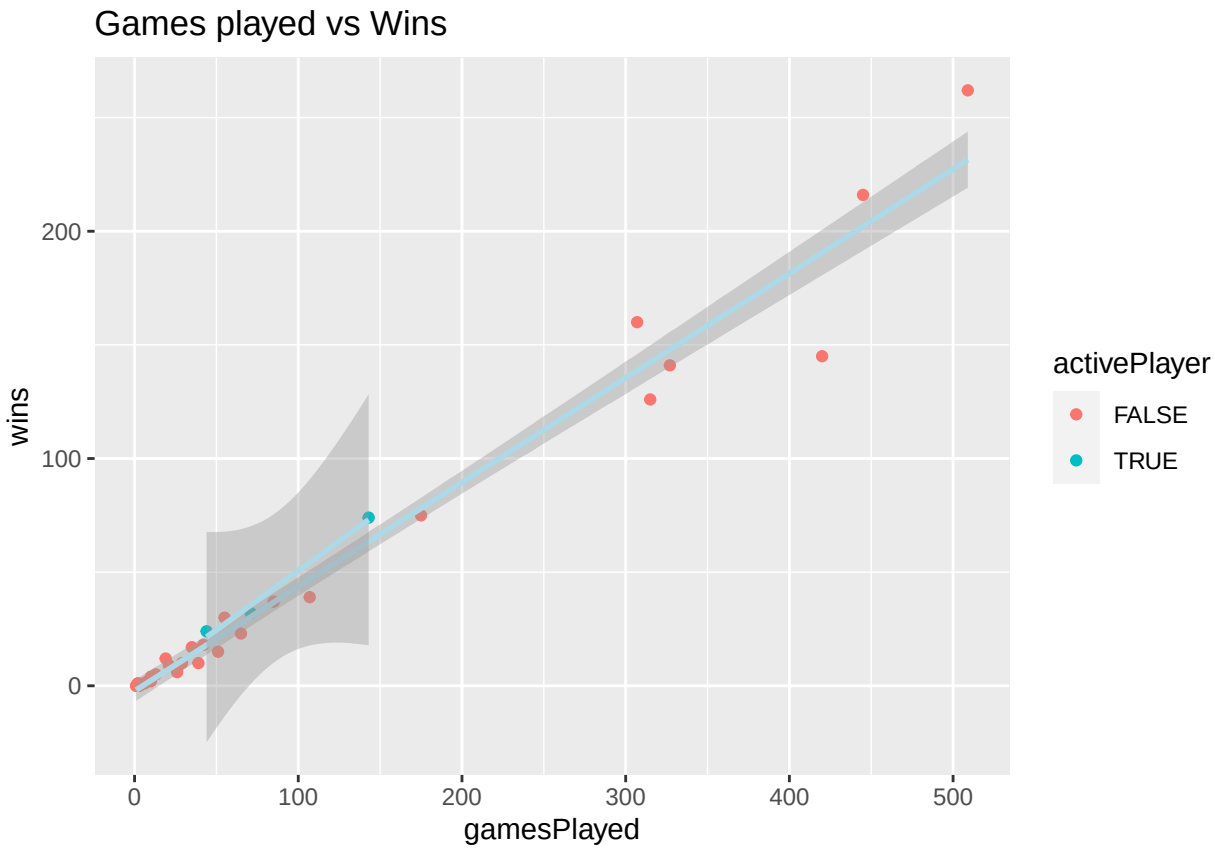
```
#scalter plot
plot(golietable15$gamesPlayed, golietable15$wins/golietable15$gamesPlayed , col="blue",
     xlab="Games Played ",
     ylab = "Games won",
     main = "Games played vs win rate")
```



The active players has larger width, which can be the effect of very lower number of players .

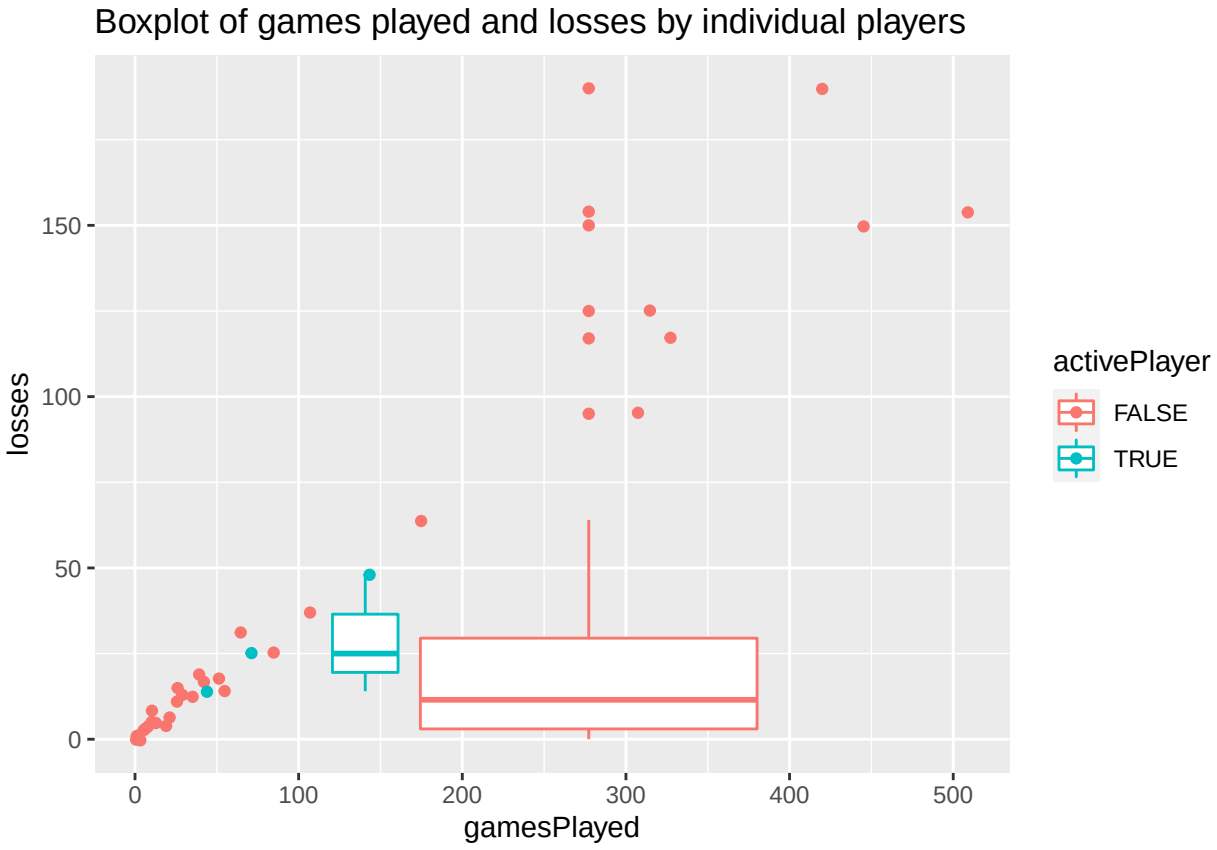
```
ggplot (golietable15, aes(x=gamesPlayed, y=wins, group=activePlayer)) + geom_point(aes(color= activePla
```

```
## `geom_smooth()` using formula 'y ~ x'
```



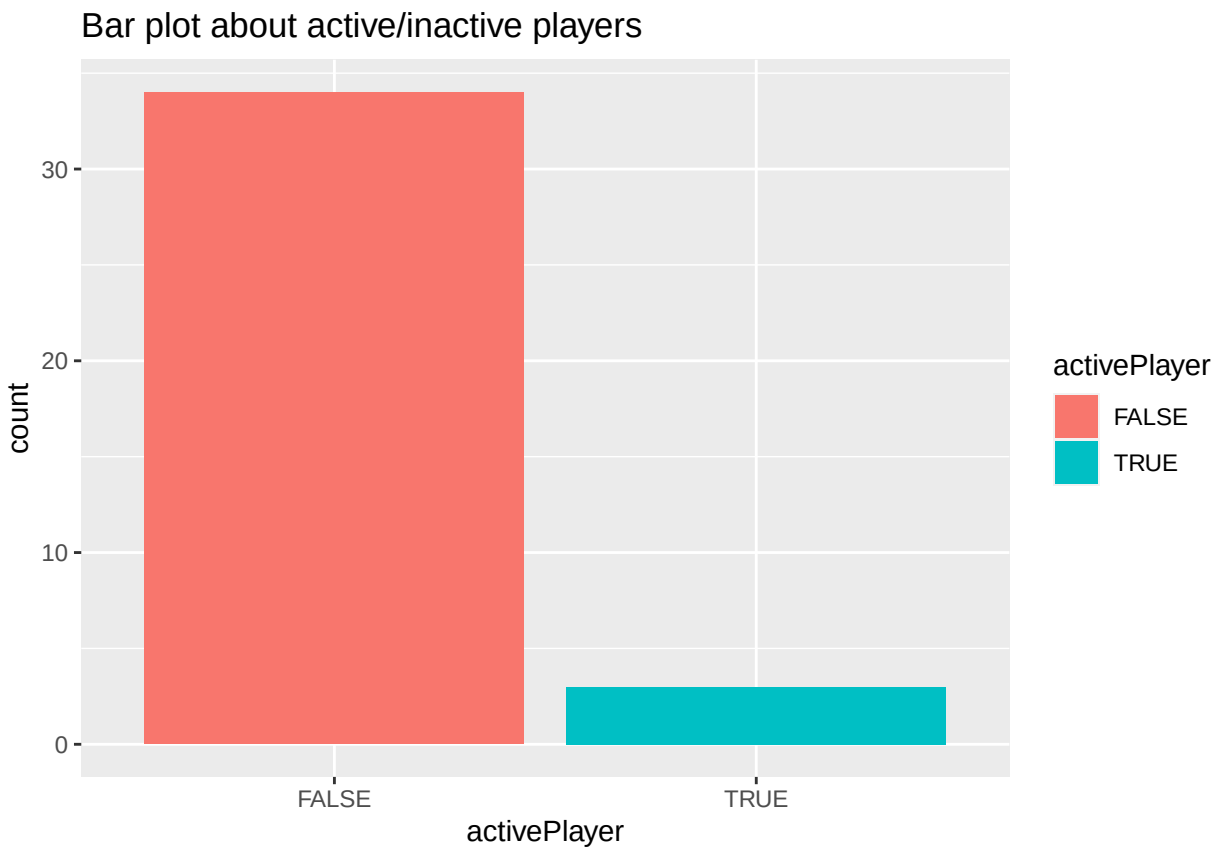
Box plot The players not tagged as active players looks to have loose the higest number of games. But in the median number of games lost is higher for active players.

```
#box plot
bxPlot1<- ggplot(data= golietable15, aes(x=gamesPlayed, y= losses, group=activePlayer, color=activePlay
bxPlot1 + geom_boxplot() + labs(title="Boxplot of games played and losses by individual players") + geo
```



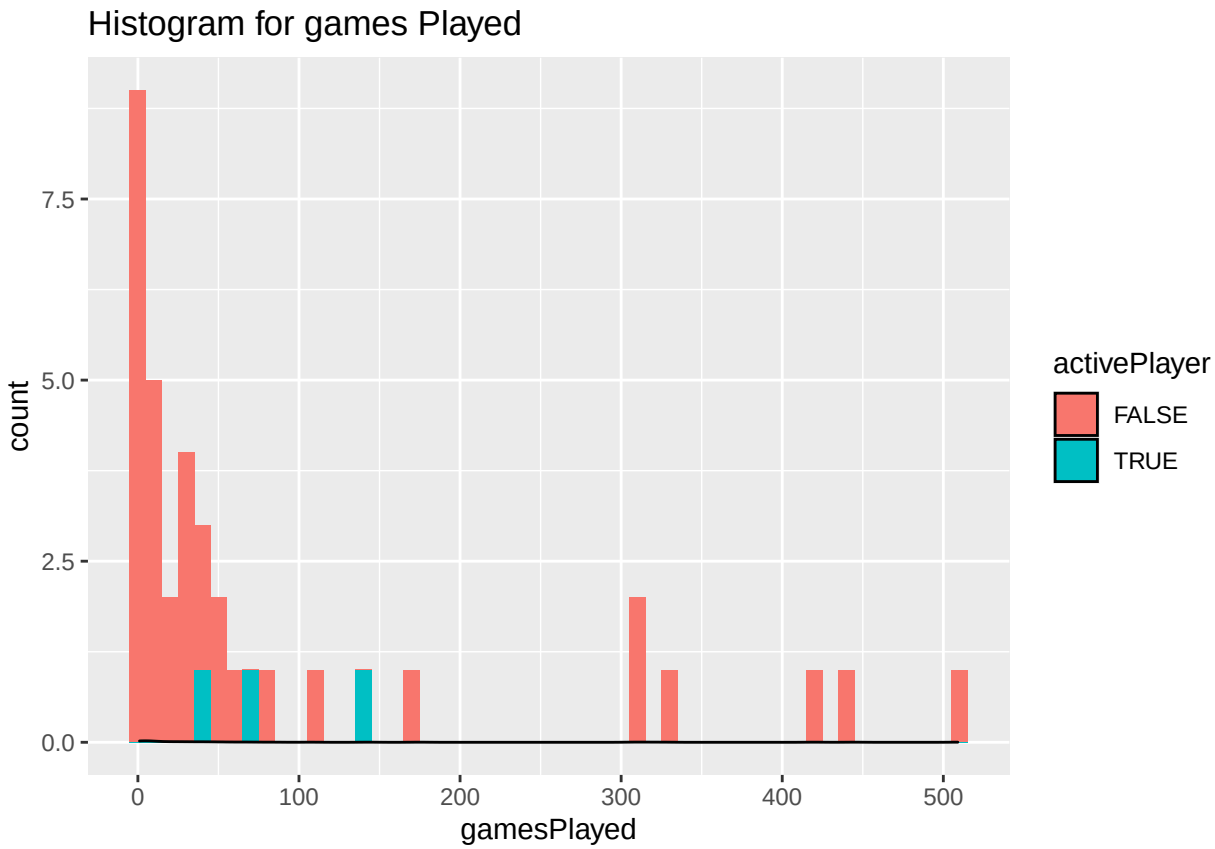
Bar plots Again, there are very few, less than 5 players are categorized as active player, which seems unreal.

```
type2<-  golietable15 %>% select(playerName, gamesPlayed, wins, activePlayer, gameTypeId)
barPlot1<- ggplot(data=type2, aes(x=activePlayer))
barPlot1 + geom_bar(aes(fill= activePlayer), position = "dodge") + labs(title = "Bar plot about active/
```



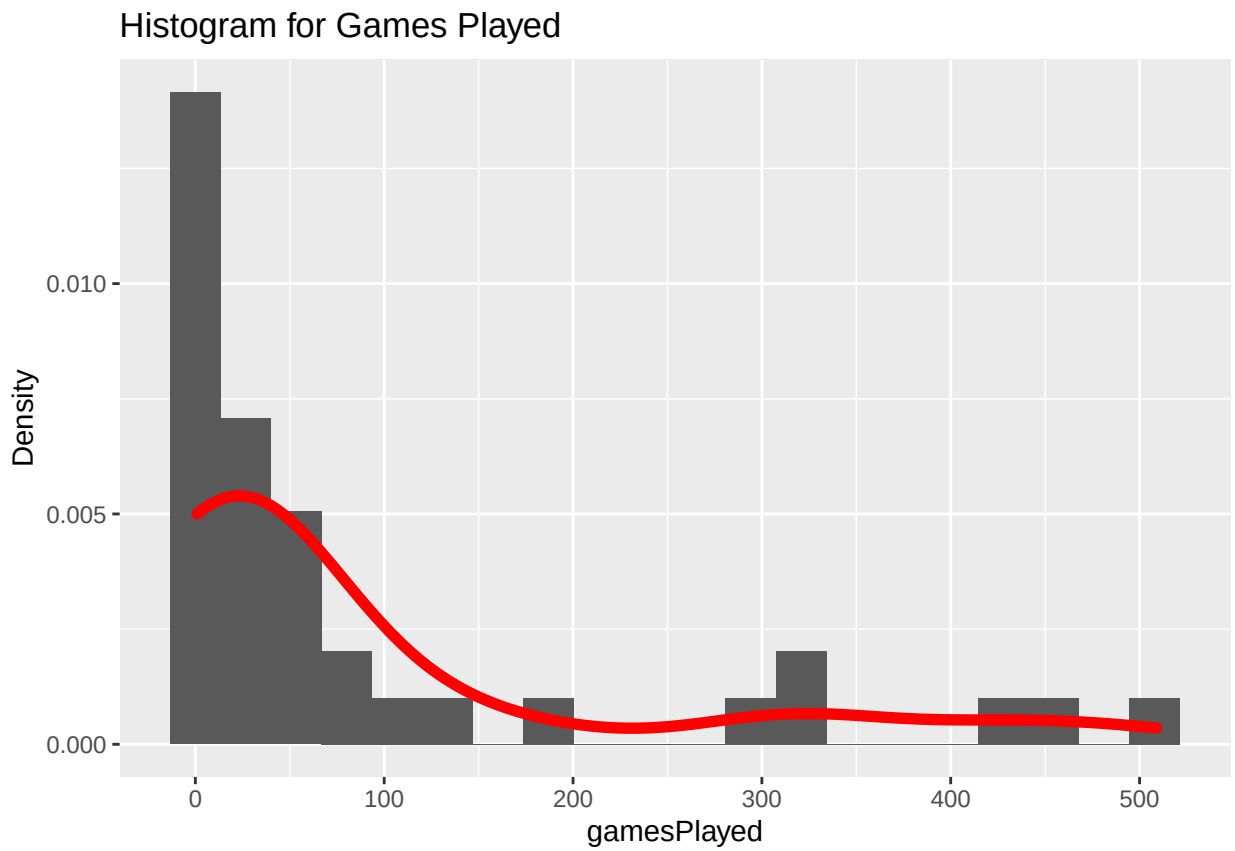
Histogram The histogram for games played grahp shows that there are very few players playing the regular season who are tagged as active player. They seem to be played between 40 to 150 games throught their entire career for the data collected period. Also, it shwos there are large group of players who played less than 5 games.the different color is contribution of two categorized playes as active or not.The distribution is extremely right skwed

```
type3<- golietable15 %>% select(wins, losses, activePlayer, gamesPlayed) #(gameTypeId==3,
histogram1<- ggplot(data=type3, aes(x=gamesPlayed))
histogram1 + geom_histogram(binwidth = 10, aes(fill= activePlayer)) + labs(title="Histogram for games P
  geom_density(adjust= 0.25, alpha=0.05)
```



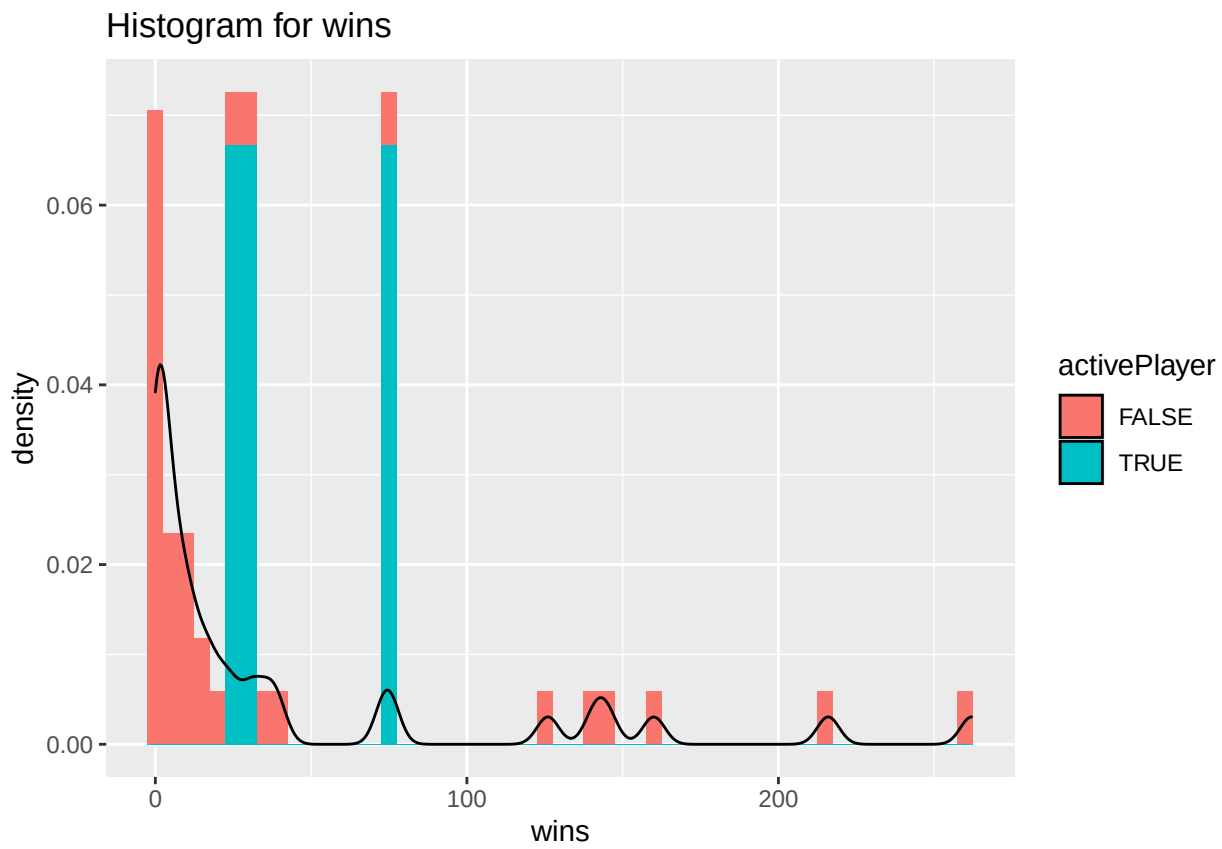
The histogram looks extremly right skewed that the players playing higher number of games are lesser

```
ggplot(golietable15, aes(x=gamesPlayed, ..density..)) + geom_histogram(bins=20) + ggtitle("Histogram fo
```



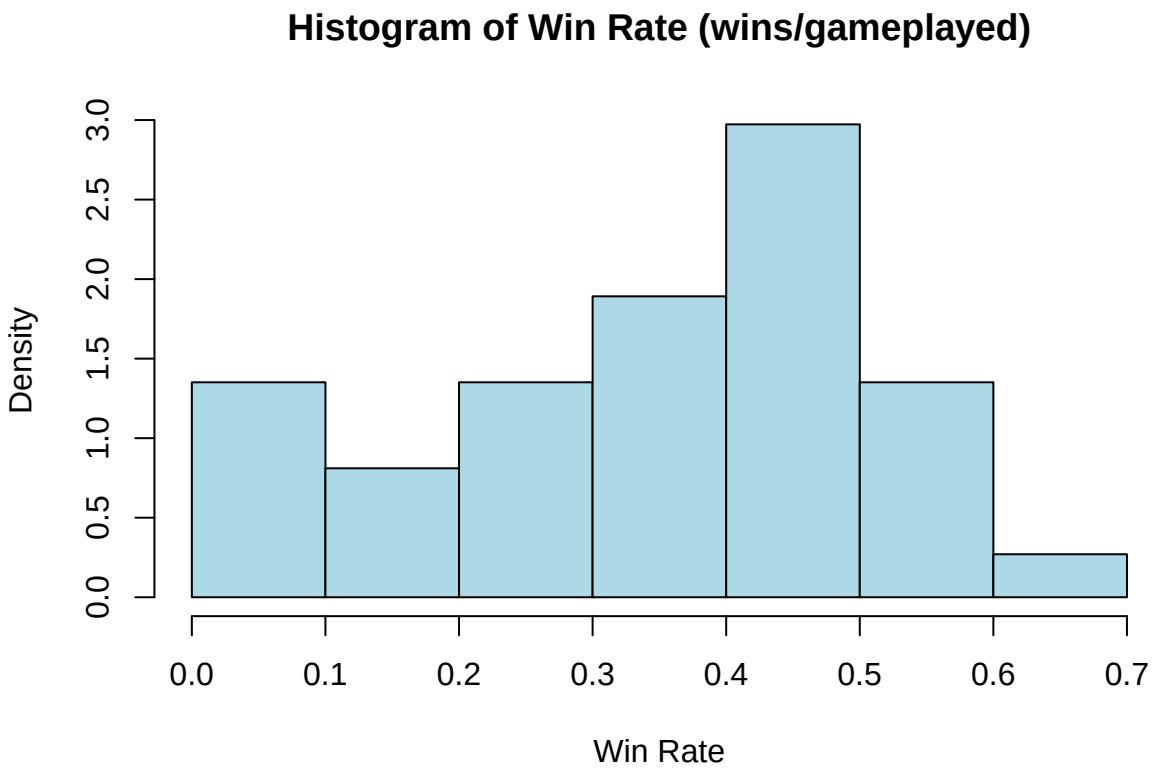
Again, the different color is contribution of two categorized playes as active or not.

```
type3<- golietable15 %>% select(wins, losses, activePlayer, gamesPlayed, playerId) #(gameTypeId==3, div
histogram1<- ggplot(data=type3, aes(x=wins))
histogram1 + geom_histogram(binwidth = 5, aes(y=..density.., fill= activePlayer )) + labs(title="Histog
```

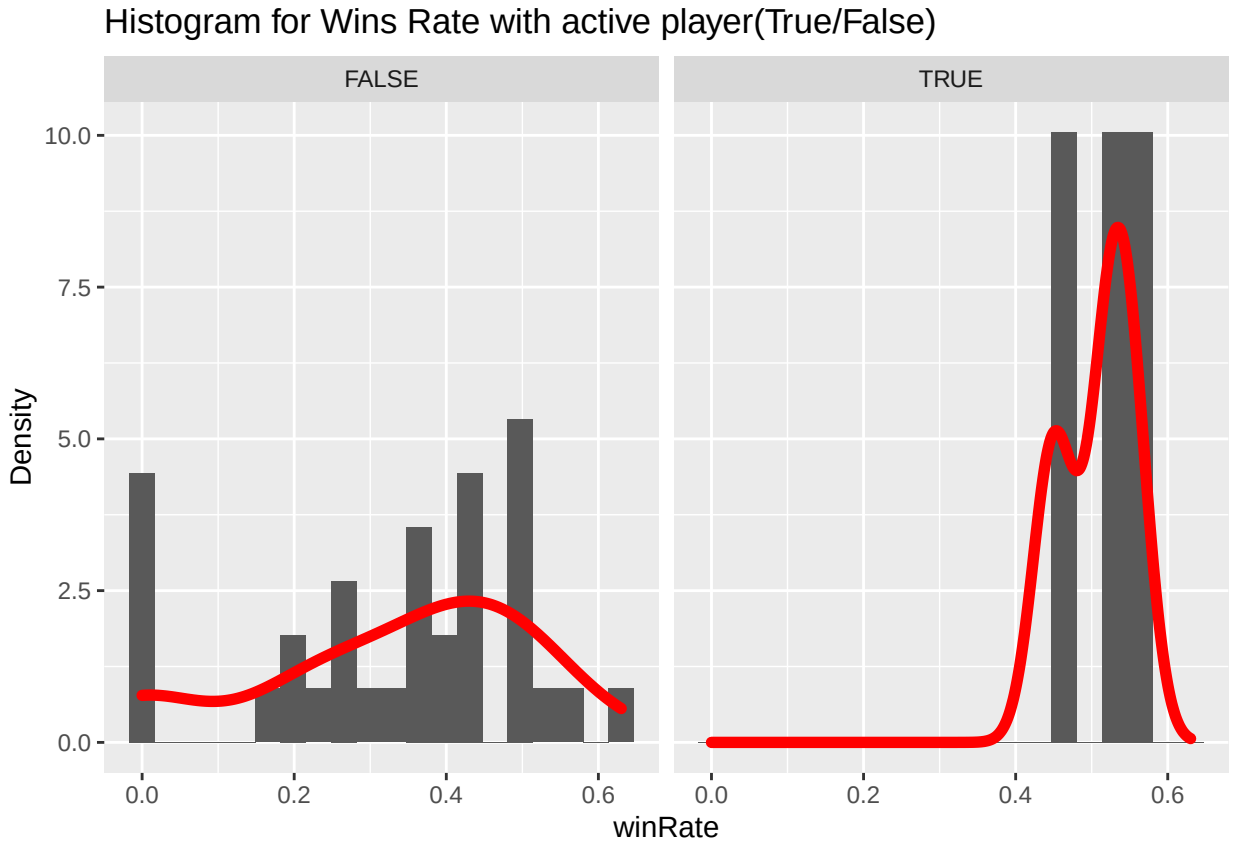


For the shake of try, the above histograms does not look preety well distributed data, so I tried below to see it on win rates, a calculated variable. The win rate is left skewed.

```
hist(wlRate$winRate, probability = T, col = "light blue", xlab = "Win Rate", main = "Histogram of Win Ra
```




```
ggplot (wlRate, aes(x=winRate, ..density..)) + geom_histogram(bins=20) + facet_wrap(golietable15$activel
```


Histogram for Wins Rate with active player(True/False)
FALSE
TRUE
Density
10.0
7.5
5.0
2.5
0.0
0.0
0.2
0.4
0.6
winRate