Midterm Project Report

Jianhao Yan, Guangyan Yu, Xuan Zhu, Megha Pandit

10/18/2018

Prepare the Dataset

The goal of this project is to explore the relationship between weather conditions and attendance at Red Sox and Celtics games in the greater Boston area. Thus, we need to prepare three datasets: one for the weather records, and the other two for adttendance at Red Sox and Celtics games, separately.

Weather data is easy to find, as the .csv file can be directly downloaded.We go to https://www.ncdc.noaa.gov/cdo-web/, then click "Search Tool". We select the "Daily Summaries" dataset from 2012-01-01~2017-12-31, with the "BOSTON, MA US" station, and we choose to collect data on Precipitation, Air Temperature, Wind, and Weather Type.

For Celtics, the thing becomes a little complicated. We used the following procedures to acquire data from multiple webpages.

Step 1. Go to https://www.basketball-reference.com/. Browse the website, and on the landing page, click on "Schedule & Results" under the section "Seasons". The reason why we choose this site over other competitors is that its urls are written in a friendly way.

As the most recent season is 17-18 (Season 18-19 will begin soon!), the page now automatically shows games in OCT,2017.

Explore all the webpages that contain the info we want. There's a naming rule for their urls:

"www.basketball-reference.com/leagues/NBA_year_games-_month.html"

So we have designed a way to scrape data from multiple web pages quickly.

Step 2. Create a table called 'urls'. 'Urls' has 9 * 6 rows (9 months per season, 6 years in total), corresponding to 54 different webpages. Fill the table with links.

Step 3. Use the table to scrape data in an easy for-loop.

Step 4. Clean and organize the data so that it is ready for analysis.

```
library(rvest)
```

```
## Loading required package: xml2
```

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
## x purrr::pluck()          masks rvest::pluck()
```

```r
#create a table to store Urls
urls <- matrix(c(1:54),54,1)


for( i in 2013:2018) {  # change the part of the year in urls
#use i in the string
  i <- as.character(i)

#create urls by the rule
OCT <- paste("https://www.basketball-reference.com/leagues/NBA",i,"games-october.html", sep = "_")
NOV <- paste("https://www.basketball-reference.com/leagues/NBA",i,"games-november.html", sep = "_")
DEC <- paste("https://www.basketball-reference.com/leagues/NBA",i,"games-december.html", sep = "_")
JAN <- paste("https://www.basketball-reference.com/leagues/NBA",i,"games-january.html", sep = "_")
FEB <- paste("https://www.basketball-reference.com/leagues/NBA",i,"games-february.html", sep = "_")
MAR <- paste("https://www.basketball-reference.com/leagues/NBA",i,"games-march.html", sep = "_")
APR <- paste("https://www.basketball-reference.com/leagues/NBA",i,"games-april.html", sep = "_")
MAY <- paste("https://www.basketball-reference.com/leagues/NBA",i,"games-may.html", sep = "_")
JUN <- paste("https://www.basketball-reference.com/leagues/NBA",i,"games-june.html", sep = "_")

i <- as.numeric(i)

#assign the links to our table

urls[i-2012] <- OCT
urls[i-2012+6] <- NOV
urls[i-2012+12] <- DEC
urls[i-2012+18] <- JAN
urls[i-2012+24] <- FEB
urls[i-2012+30] <- MAR
urls[i-2012+36] <- APR
urls[i-2012+42] <- MAY
urls[i-2012+48] <- JUN
}

# 'Urls' has been created successfully.
# Now Scrape data from the website by using our table
for (i in 1:54){
templink <- read_html(urls[i])
a <- templink%>%html_nodes("table")%>%.[[1]]%>%html_table()

#data transformation
colnames(a)[5] <- "City"
colnames(a)[4] <- "3"
colnames(a)[7] <- "1"
colnames(a)[8] <- "2"
a <- a %>%
select(Date,City,Attend.)%>%
filter(City =="Boston Celtics")

if (i == 1) {
  Celtics <- a
}
#conbime all the data into a single table
```

```
Celtics <- rbind(a,Celtics)
}

Celtics$Date <- substr(Celtics$Date,6,17)
Celtics$Date <- as.Date(Celtics$Date,format='%B %d, %Y')
```

For Red Sox games, the procedures are similar to Celtics.

Step 1. Go to https://www.baseball-reference.com/teams/BOS/2017-schedule-scores.shtml. In this website, there is a table contains data of mathes including date, attendance and wheather it is a home or away game, etc.

Step 2. We find that with different game year, the URL of this website just changes in the year number, so that it is not hard for us to gather the game-by-game table from 2012 to 2017. The procedure of scraping data is an easy for loop.

Step 3. Clean and organize the data so that it is ready for analysis.

```
#baseball data
library(XML)
```

```
##
## Attaching package: 'XML'
```

```
## The following object is masked from 'package:rvest':
##
##     xml
```

```
library(RCurl)
```

```
## Loading required package: bitops
```

```
##
## Attaching package: 'RCurl'
```

```
## The following object is masked from 'package:tidyr':
##
##     complete
```

```
library(rlist)
library(stringr)
library(readxl)

weather <- read.csv("weather.csv")

url1 <- "https://www.baseball-reference.com/teams/BOS/"
url2 <- "-schedule-scores.shtml"

years <- c(2012:2017)

urls2 <- str_c(url1, years, url2, sep = "")

filenames <- str_c("mr", years, sep = "")

N <- length(urls2)

for (i in 1:N){
  suppressMessages(
    assign(filenames[i], readHTMLTable(getURL(urls2[i])))
```

```
  )

  file <- get(filenames[i])[[1]]
  file<-file[!str_detect(file$`Gm#`,"Gm#"),]
  colnames(file)[1]<-"YYYY"
  file[,1]<-years[i]
  if(i == 1){
    redsox <- file
  }

  else{
    redsox <- rbind.data.frame(redsox, file)
  }
}

redsox$Date<-paste(redsox$Date,redsox$YYYY,sep = ",")
redsox$Date<-str_replace_all(redsox$Date," \\((.*?)\\)","")
redsox$Date<-as.Date(redsox$Date, format="%A, %b %d, %Y")
redsox<-redsox[!str_detect(redsox[,5],"@"),]

redsox_attendence<-subset(redsox,select=c(Date,Time,Attendance))
str(redsox_attendence)#Date:Date, Time:factor, Attendence:factor
```

```
## 'data.frame':    486 obs. of  3 variables:
##  $ Date      : Date, format: "2012-04-13" "2012-04-14" ...
##  $ Time      : Factor w/ 153 levels "2:13","2:18",..: 49 48 42 38 54 33 48 67 53 60 ...
##  $ Attendance: Factor w/ 938 levels "11,502","11,722",..: 56 121 121 126 129 120 54 115 86 69 ...
```

```
redsox_attendence$Attendance<-str_replace_all(redsox_attendence$Attendance,",","")
redsox_attendence$Attendance<-as.numeric(redsox_attendence$Attendance)
```

Now all the datasets are ready for analysis.

EDA

(A) Red Sox

Before we formally produce EDA, it is always good to take a look at the dataset first, and then think about the problem we are trying to throw out. As we can see from the weather table, there are so many different types of weather conditions, such as fog, snow and wind. Faced with multiple weather types, our group tends to analyse the influence in a simple way. So we define the date with special weather conditions as abnormal weather.

```
str(redsox)
```

```
## 'data.frame':    486 obs. of  21 variables:
##  $ YYYY          : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
##  $ Date          : Date, format: "2012-04-13" "2012-04-14" ...
##  $               : Factor w/ 2 levels "","boxscore": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Tm            : Factor w/ 2 levels "BOS","Tm": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Ã             : Factor w/ 2 levels "","@": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Opp           : Factor w/ 30 levels "ATL","BAL","CHC",..: 16 16 16 16 17 17 11 11 12 12 ...
##  $ W/L           : Factor w/ 5 levels "L","L-wo","W",..: 3 3 3 1 1 1 1 1 3 1 ...
##  $ R             : Factor w/ 21 levels "0","1","10","11",..: 5 6 13 1 10 10 9 16 4 10 ...
```

```
##  $ RA             : Factor w/ 21 levels "0","1","10","13",..: 8 12 11 2 7 13 13 6 13 12 ...
##  $ Inn            : Factor w/ 13 levels "","10","11","12",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ W-L            : Factor w/ 843 levels "0-1","0-2","0-3",..: 28 50 65 66 67 68 69 64 9 10 ...
##  $ Rank           : Factor w/ 6 levels "3","4","5","Rank",..: 3 3 2 3 3 3 3 3 3 3 ...
##  $ GB             : Factor w/ 73 levels "1.0","10.0","10.5",..: 18 18 1 18 29 29 30 32 30 32 ...
##  $ Win            : Factor w/ 287 levels "Aceves","Albers",..: 12 16 72 83 55 42 70 84 16 73 ...
##  $ Loss           : Factor w/ 329 levels "Aceves","Alvarez",..: 63 6 51 8 40 10 13 1 49 24 ...
##  $ Save           : Factor w/ 99 levels "","Aceves","Bailey",..: 1 1 2 25 1 1 1 1 1 19 ...
##  $ Time           : Factor w/ 153 levels "2:13","2:18",..: 49 48 42 38 54 33 48 67 53 60 ...
##  $ D/N            : Factor w/ 3 levels "D","D/N","N": 1 1 1 1 3 3 1 1 3 3 ...
##  $ Attendance     : Factor w/ 938 levels "11,502","11,722",..: 56 121 121 126 129 120 54 115 86 69 .
##  $ Streak         : Factor w/ 22 levels "-","--","---",..: 9 10 11 1 2 3 4 5 9 1 ...
##  $ Orig. Scheduled: Factor w/ 14 levels "","2012-04-22 (Rain)",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
str(redsox_attendence)
```

```
## 'data.frame':    486 obs. of  3 variables:
##  $ Date      : Date, format: "2012-04-13" "2012-04-14" ...
##  $ Time      : Factor w/ 153 levels "2:13","2:18",..: 49 48 42 38 54 33 48 67 53 60 ...
##  $ Attendance: num  37032 38024 38024 38108 38229 ...
```

```
str(weather)
```

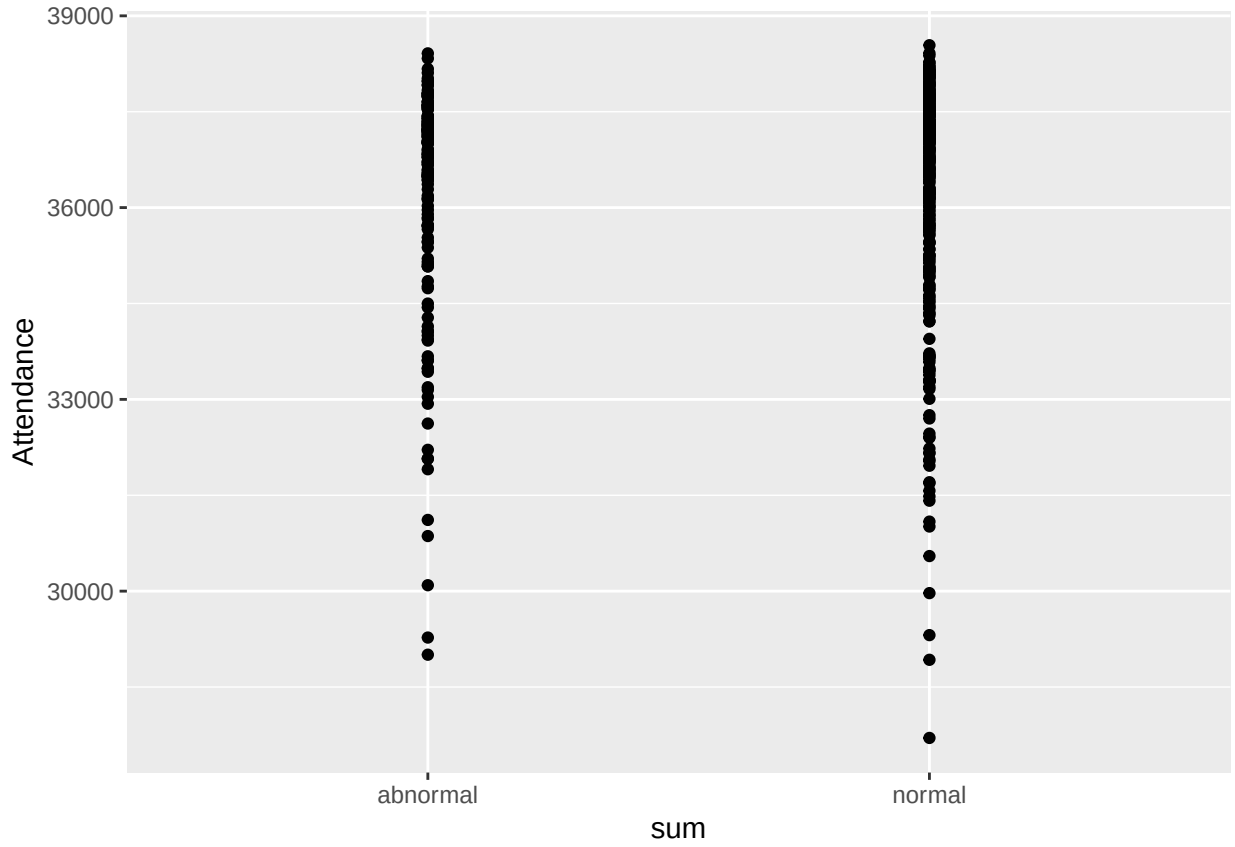
```
## 'data.frame':    2192 obs. of  52 variables:
##  $ STATION        : Factor w/ 1 level "USW00014739": 1 1 1 1 1 1 1 1 1 1 ...
##  $ NAME           : Factor w/ 1 level "BOSTON, MA US": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LATITUDE       : num  42.4 42.4 42.4 42.4 42.4 ...
##  $ LONGITUDE      : num  -71 -71 -71 -71 -71 ...
##  $ ELEVATION      : num  3.7 3.7 3.7 3.7 3.7 3.7 3.7 3.7 3.7 3.7 ...
##  $ DATE           : Factor w/ 2192 levels "2012-01-01","2012-01-02",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ AWND           : num  9.17 13.87 14.76 11.86 11.41 ...
##  $ AWND_ATTRIBUTES: Factor w/ 2 levels ",,W",",,X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ PRCP           : num  0.01 0.01 0 0 0 0 0 0 0 0.02 ...
##  $ PRCP_ATTRIBUTES: Factor w/ 4 levels ",,W,2400",",,X,2400",..: 2 2 2 2 2 4 2 2 2 2 ...
##  $ SNOW           : num  0 0 0 0 0 0 0 0 0 0.5 ...
##  $ SNOW_ATTRIBUTES: Factor w/ 5 levels ",,W",",,X","T,,W",..: 2 2 2 2 2 4 2 2 2 2 ...
##  $ SNWD           : logi  NA NA NA NA NA NA ...
##  $ SNWD_ATTRIBUTES: logi  NA NA NA NA NA NA ...
##  $ TAVG           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TAVG_ATTRIBUTES: Factor w/ 2 levels "","H,,S": 1 1 1 1 1 1 1 1 1 1 ...
##  $ TMAX           : int  52 50 35 28 39 48 60 45 40 47 ...
##  $ TMAX_ATTRIBUTES: Factor w/ 2 levels ",,W",",,X": 2 2 2 2 2 2 2 2 2 2 ...
##  $ TMIN           : int  39 34 14 10 25 28 30 30 25 30 ...
##  $ TMIN_ATTRIBUTES: Factor w/ 2 levels ",,W",",,X": 2 2 2 2 2 2 2 2 2 2 ...
##  $ WT01           : int  NA NA NA NA NA NA NA NA NA 1 ...
##  $ WT01_ATTRIBUTES: Factor w/ 3 levels "",",,W",",,X": 1 1 1 1 1 1 1 1 1 2 ...
##  $ WT02           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ WT02_ATTRIBUTES: Factor w/ 3 levels "",",,W",",,X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ WT03           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ WT03_ATTRIBUTES: Factor w/ 3 levels "",",,W",",,X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ WT04           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ WT04_ATTRIBUTES: Factor w/ 3 levels "",",,W",",,X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ WT05           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ WT05_ATTRIBUTES: Factor w/ 3 levels "",",,W",",,X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ WT06           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ WT06_ATTRIBUTES: Factor w/ 2 levels "",",,W": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ WT08           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ WT08_ATTRIBUTES: Factor w/ 3 levels "","",,W","",,X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ WT09           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ WT09_ATTRIBUTES: Factor w/ 3 levels "","",,W","",,X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ WT13           : int  NA NA NA NA NA NA 1 NA NA 1 ...
##  $ WT13_ATTRIBUTES: Factor w/ 2 levels "","",,X": 1 1 1 1 1 1 2 1 1 2 ...
##  $ WT14           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ WT14_ATTRIBUTES: Factor w/ 2 levels "","",,X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ WT15           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ WT15_ATTRIBUTES: Factor w/ 2 levels "","",,X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ WT16           : int  1 1 NA NA NA NA NA NA NA NA ...
##  $ WT16_ATTRIBUTES: Factor w/ 2 levels "","",,X": 2 2 1 1 1 1 1 1 1 1 ...
##  $ WT17           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ WT17_ATTRIBUTES: Factor w/ 2 levels "","",,X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ WT18           : int  NA NA NA NA NA 1 NA NA NA 1 ...
##  $ WT18_ATTRIBUTES: Factor w/ 2 levels "","",,X": 1 1 1 1 1 2 1 1 1 2 ...
##  $ WT19           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ WT19_ATTRIBUTES: Factor w/ 2 levels "","",,X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ WT22           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ WT22_ATTRIBUTES: Factor w/ 2 levels "","",,X": 1 1 1 1 1 1 1 1 1 1 ...
```

```
type_code = c("WT01","WT02","WT03","WT04","WT05","WT06","WT08","WT09","WT13",
              "WT14","WT15","WT16","WT17","WT18","WT19","WT22")
```

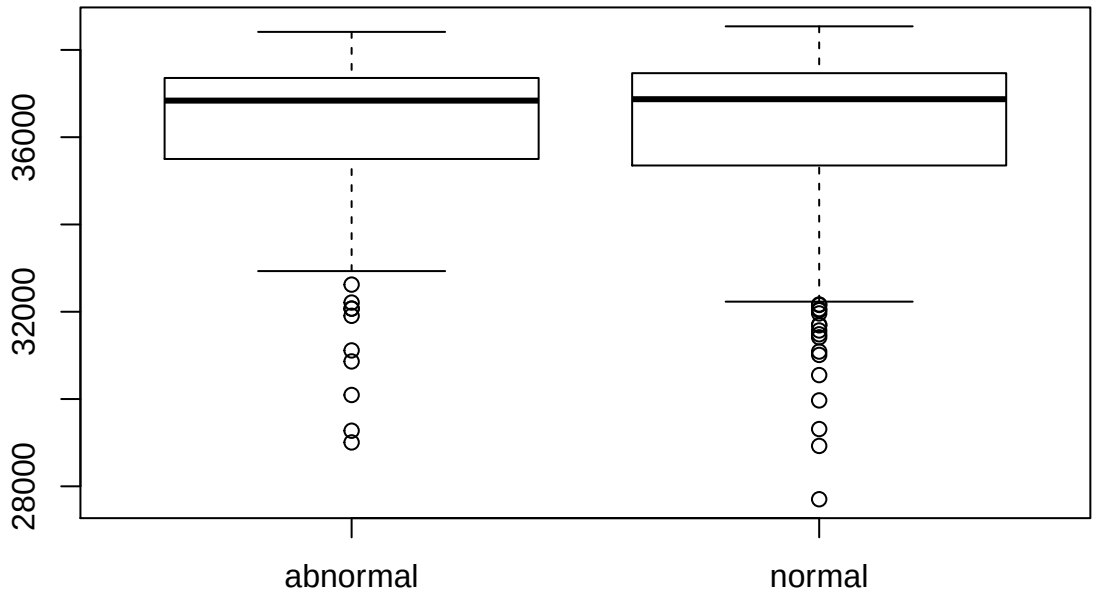
When trying to do the scatter plot of the weather conditions versus attendace number, we cannot see any clear relationship between them. It is only possible to conclude that the maximum is between 36000 and 39000 for both normal weather and abnormal weather.

```
library(dplyr)
library(ggplot2)
weather1 = weather[,c("WT01","WT02","WT03","WT04","WT05","WT06","WT08","WT09","WT13",
                      "WT14","WT15","WT16","WT17","WT18","WT19","WT22")]
weather1[is.na(weather1)] = 0
a = rowSums(weather1)
weather$sum = a
for (i in 1:dim(weather)[1]){
  if (as.integer(weather$sum[i]) >= 1) {
    weather$sum[i] = "abnormal"
  }
  else{
    weather$sum[i] = "normal"
  }
}
colnames(weather)[6]<-"Date"

##Join the two tables
weather$Date = as.Date(weather$Date)
table_type<-inner_join(weather,redsox_attendence,by="Date")
weather$sum = as.character(weather$sum)
weather2 <- table_type%>%
  group_by(sum)%>%
  summarise(mean_attendance=mean(Attendance))
##Plots
ggplot(data = table_type)+geom_point(aes(x=sum,y=Attendance))
```



So we switch to make the boxplot. We find that there is no difference in the average attendances between under normal weather and under abnormal weather. To verify this claim, we perform the t-test.The conculsion is that we should accept the null hypothesis that the average attendances are the same. So it makes no big difference in the attendances of Red Sox games even though weather conditions are bad.

```
boxplot(data=table_type,Attendance~sum)
```




```
##t-test
table_type2<-table_type[,c("sum","Attendance")]
table_normal<-filter(table_type2,sum=='normal')
table_abnormal<-filter(table_type2,sum=="abnormal")
t.test(table_abnormal$Attendance,table_normal$Attendance)
```

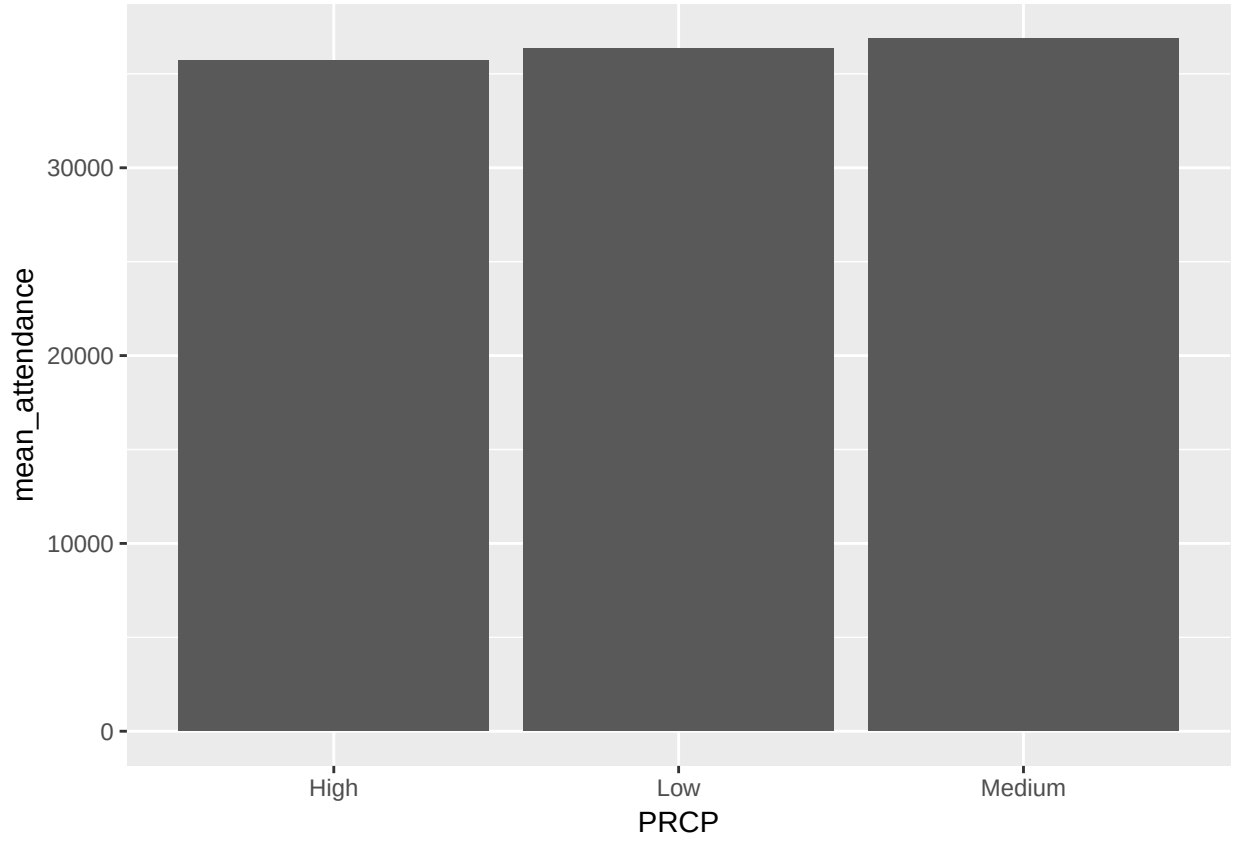
```
##
##  Welch Two Sample t-test
##
## data:  table_abnormal$Attendance and table_normal$Attendance
## t = -0.17574, df = 319.74, p-value = 0.8606
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -384.1435  321.1446
## sample estimates:
## mean of x mean of y
##  36164.53  36196.03
```

```
##no obvious difference between the attendances on normal and abnormal days.
```

Besides, we can take the influence of raining on the attendance of Red Sox games as an exmaple. First, define the rainy level as “low”, “medium” and “high”. Then we use ggplot to graph the bar of the rainy level vs average attendaces. We still do not find any differences in the average attendance among different rainy levels. The last graph also indicates that the rain really has few influences on the attendances.

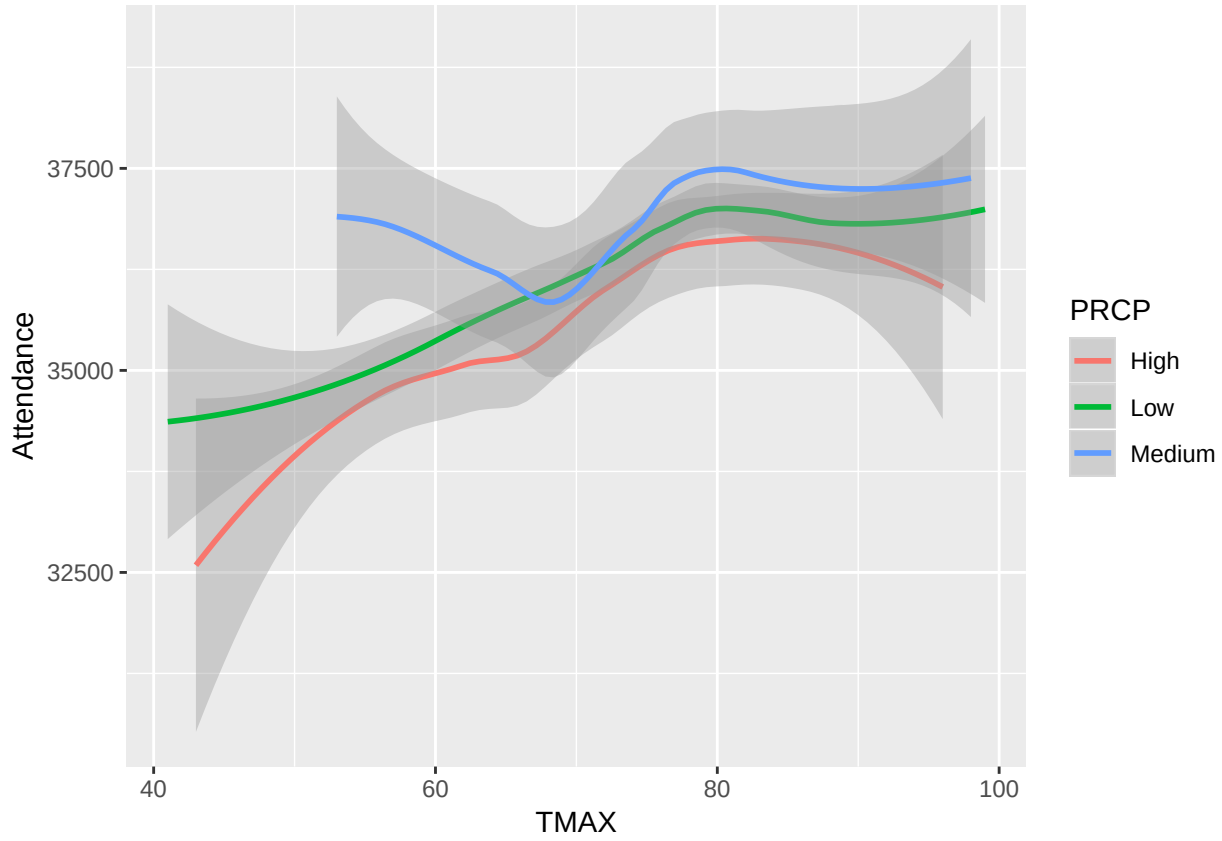
```
library(dplyr)
table_rain<-weather[,c('Date','PRCP')]
colnames(table_rain)[1]<-"Date"
```

```
table_rain$Date = as.Date(table_rain$Date)
Attendance_rain<-inner_join(redsox_attendence,table_rain,by="Date")

Attendance_rain$PRCP <- ifelse(Attendance_rain$PRCP==0,"Low",
                          ifelse(Attendance_rain$PRCP>0 & Attendance_rain$PRCP<0.02, "Medium", "High"))

Attendance_rain1<-Attendance_rain%>%
  group_by(PRCP)%>%
  summarise(mean_attendance=mean(Attendance))
p_red2<-ggplot(Attendance_rain1, aes(PRCP, mean_attendance))+geom_bar(stat = "identity")
p_red2
```




```
Attendance_rain2<-inner_join(redsox_attendence,table_rain,by="Date")
table_tmax<-weather[,c("Date","TMAX")]
table_tmax_rain<-inner_join(Attendance_rain,table_tmax,by="Date")
table_tmax_rain$PRCP<-as.factor(table_tmax_rain$PRCP)
p_red3<-ggplot(data =table_tmax_rain) +
  geom_smooth(mapping = aes(x=TMAX, y=Attendance,color=PRCP))
p_red3
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

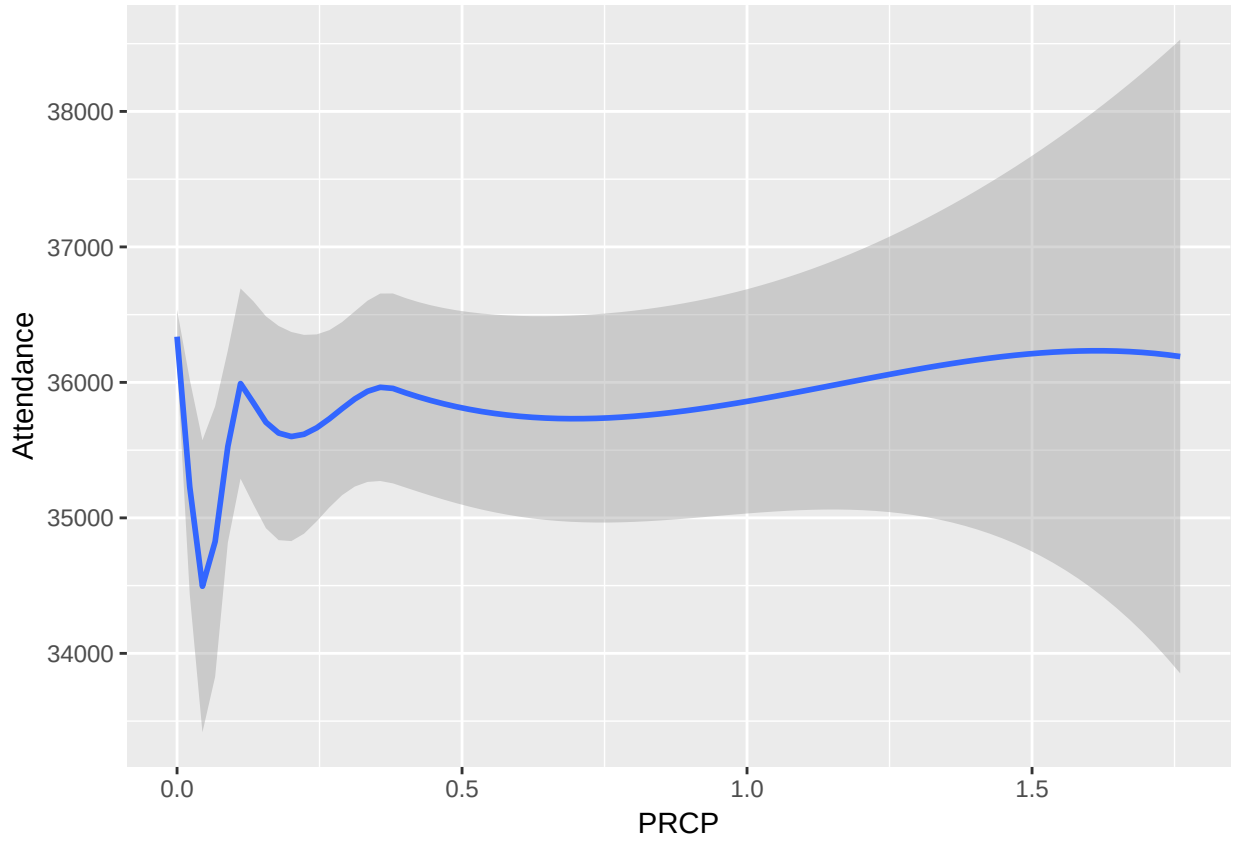



```
join_rain_redsox<-inner_join(weather,redsox_attendence,by="Date")
p_red4<-ggplot(data = join_rain_redsox)+geom_smooth(mapping = aes(x=PRCP,y=Attendance))
p_red4
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at -0.0088

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 0.0288

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 1.3804e-015

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 0.0004

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used
## at -0.0088

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 0.0288

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
```

```
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal
## condition number 1.3804e-015
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other
## near singularities as well. 0.0004
```

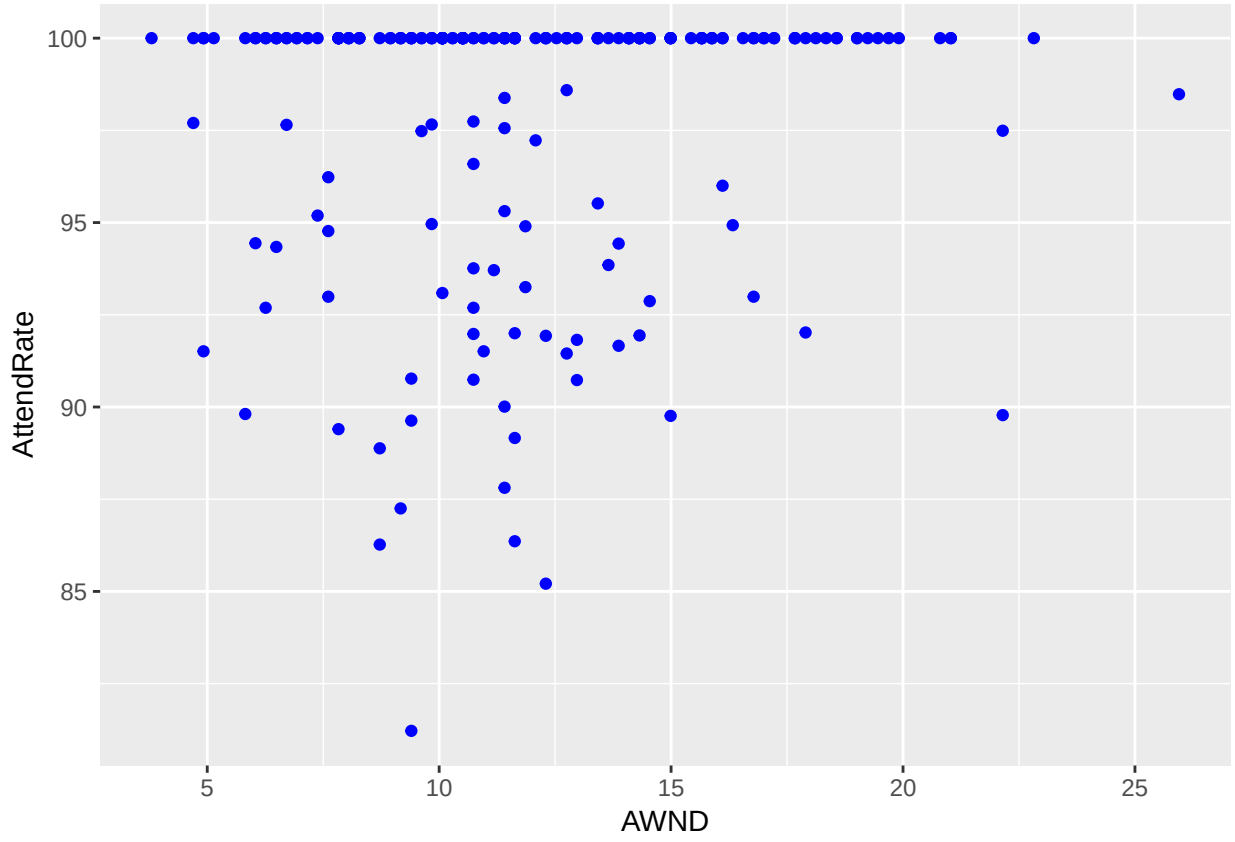


(B)Celtics

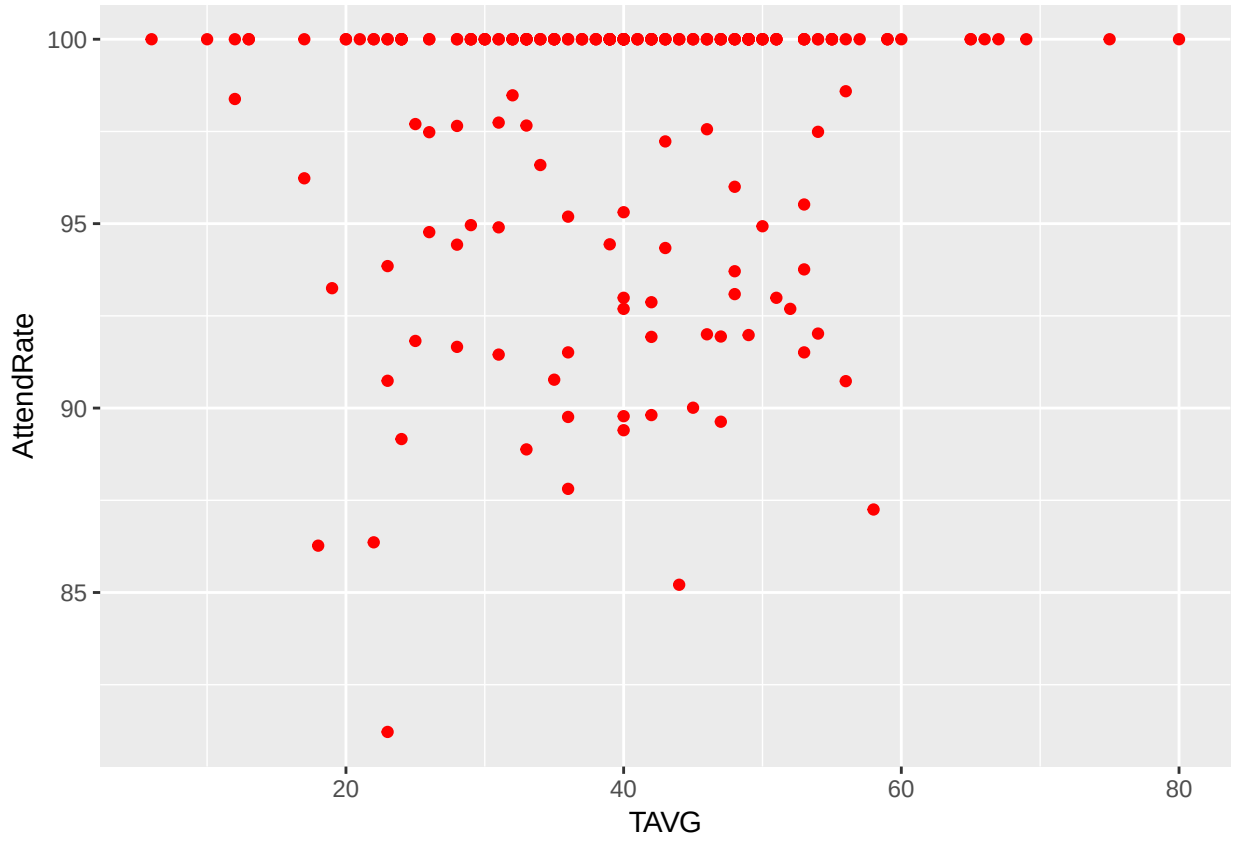
The same situation happens to Celtics games: they are not largely affected by the weather condition in general.

```
#merge two tables into one
sub.weather <-weather
colnames(sub.weather)[6] <- "Date"
sub.weather$Date <- as.Date(sub.weather$Date)
EDAtable <- left_join(Celtics,sub.weather, by = "Date")

#clear out unnecessary data and edit on existing data
EDAtable <-EDAtable%>%
select(Date,Attend.,AWND,TAVG)
EDAtable <- na.omit(EDAtable)
EDAtable$Attend. <-as.numeric(gsub(",","",EDAtable$Attend.))
AttendRate <- (EDAtable$Attend./18624) * 100
AttendRate <- round(AttendRate,2)
```

```
EDAtable <-cbind(EDAtable,AttendRate)

#now we use EDAtable to explore the relationship between attendace and average wind speed & attendance
library(ggplot2)
ggplot(data=EDAtable)+
  geom_point(aes(x=AWND,y=AttendRate),color="blue")
```

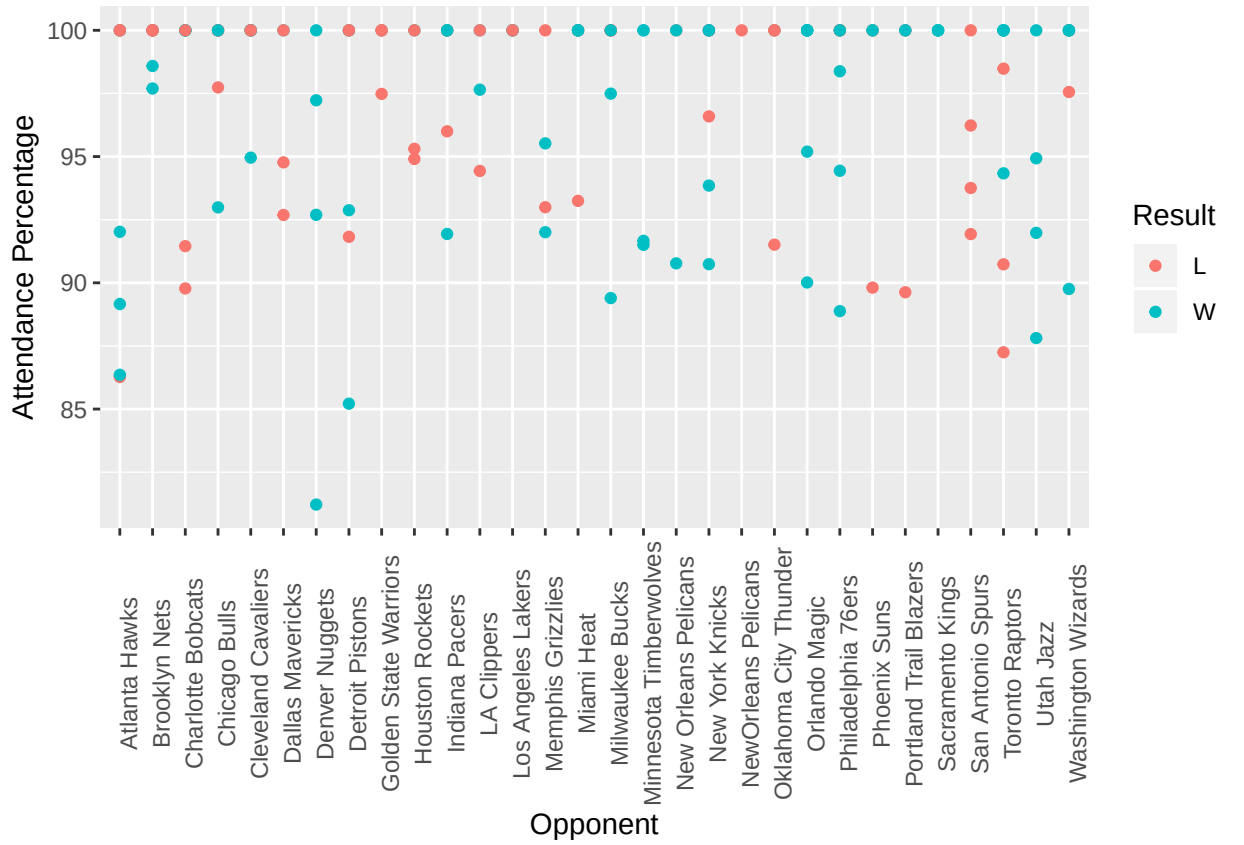



```
ggplot(data=EDAtable)+
  geom_point(aes(x=TAVG,y=AttendRate),color="red")
```



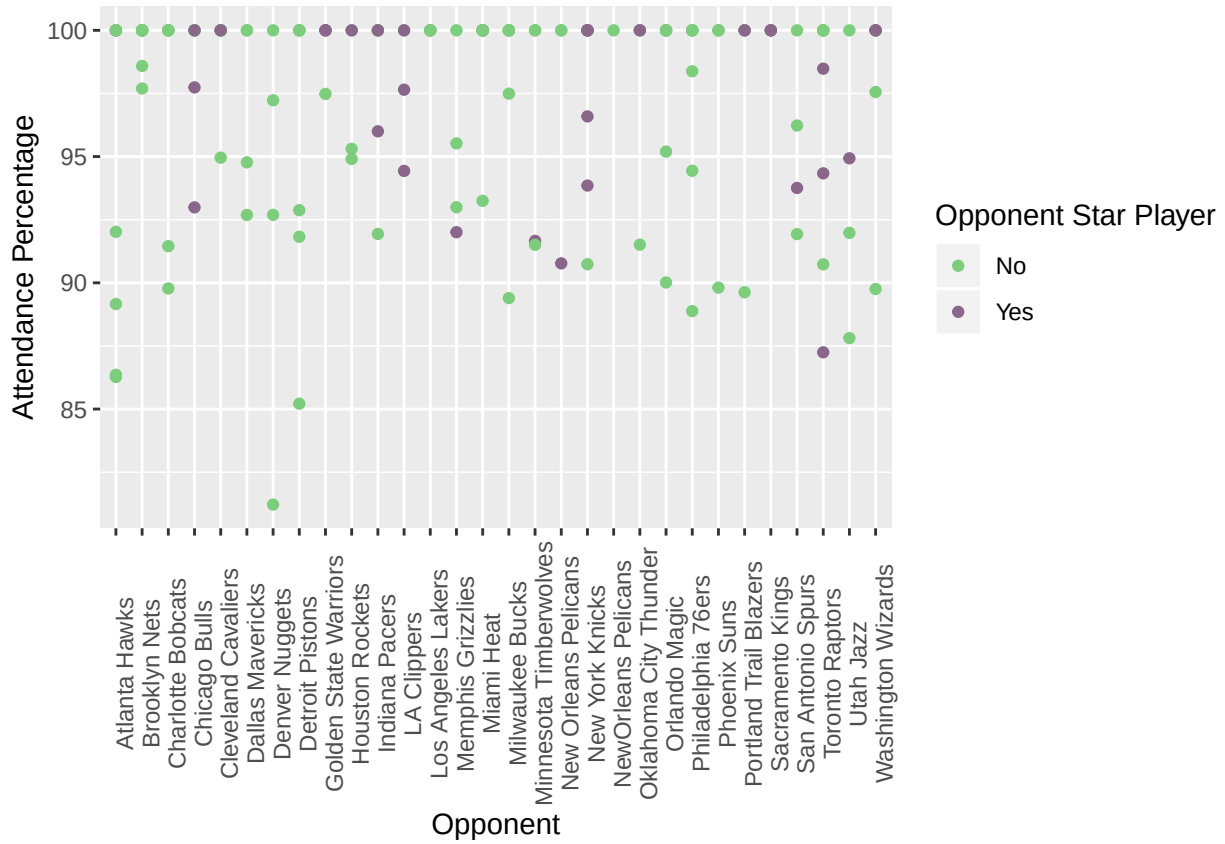
Since Basketball is an indoor game, there may be other major factors, other than weather, that could affect the spectator attendance at each game. Therefore, we explored the effect of the popularity of the opponent team on game attendance.

```
library(ggplot2)
#Collected data on opponent team from the ESPN website
library(readxl)
BCeltics <- read_excel("BCeltics.xlsx")

attd <- BCeltics$Attendance*100/18624 #taking the attendance percentage

names(BCeltics)[6] <- paste("OSP")

#plotting the attendance percentage in each of the games with the opponent listed
ggplot(BCeltics)+
  geom_point(aes(x = Opponent, y = attd, color = Result))+
  ylab("Attendance Percentage")+
  theme(axis.text.x = element_text(angle = 90))
```



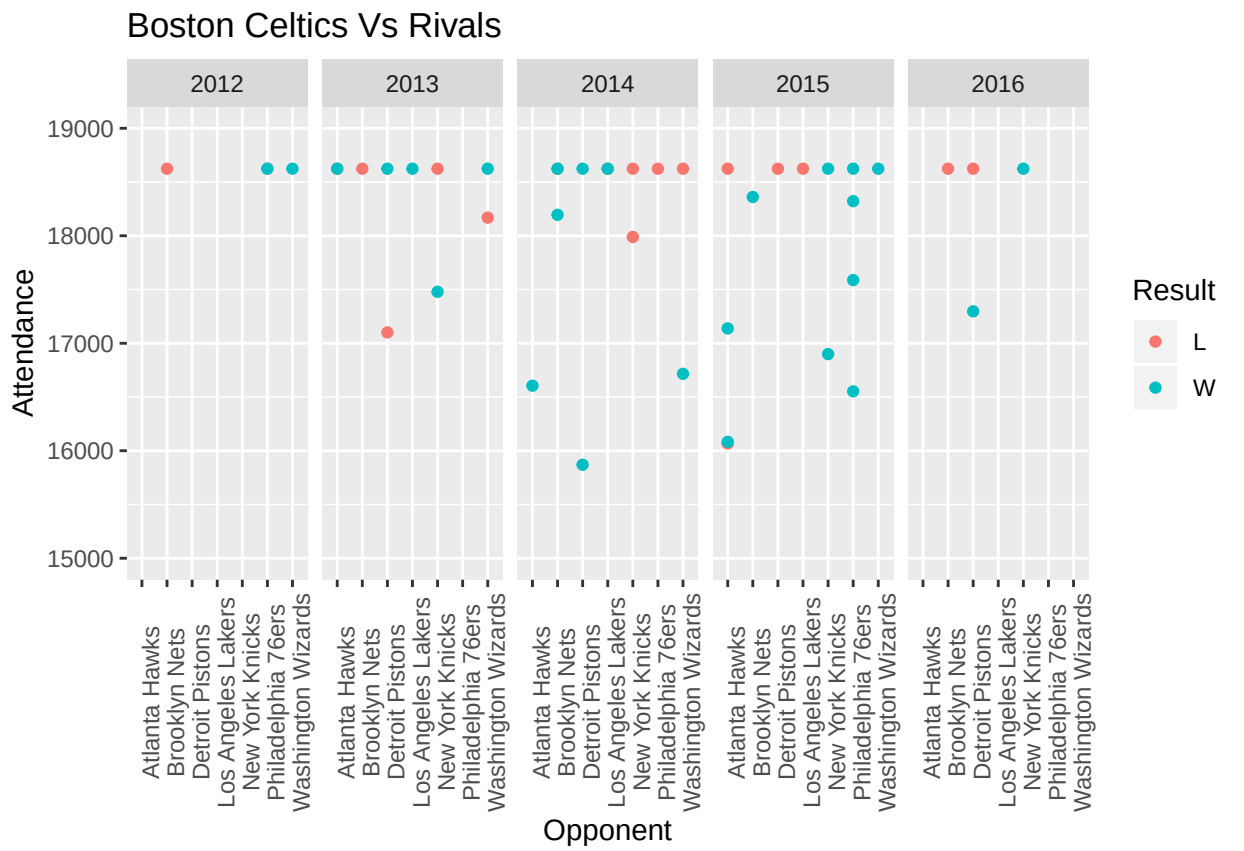
The attendance for each game has been plotted above and it shows that the overall attendance percentage is above 85%. There are small variations in the attendance that may be because of the popularity of the opponent team.

```
#checking if having a star player in the opponent team has an effect on attendance
legend_tilte <- "Opponent Star Player"
ggplot(BCeltics)+
  geom_point(aes(x = Opponent, y = attd, color = OSP))+
  ylab("Attendance Percentage")+
  theme(axis.text.x = element_text(angle = 90))+
  scale_color_manual(legend_tilte, values = c("palegreen3", "plum4"))
```

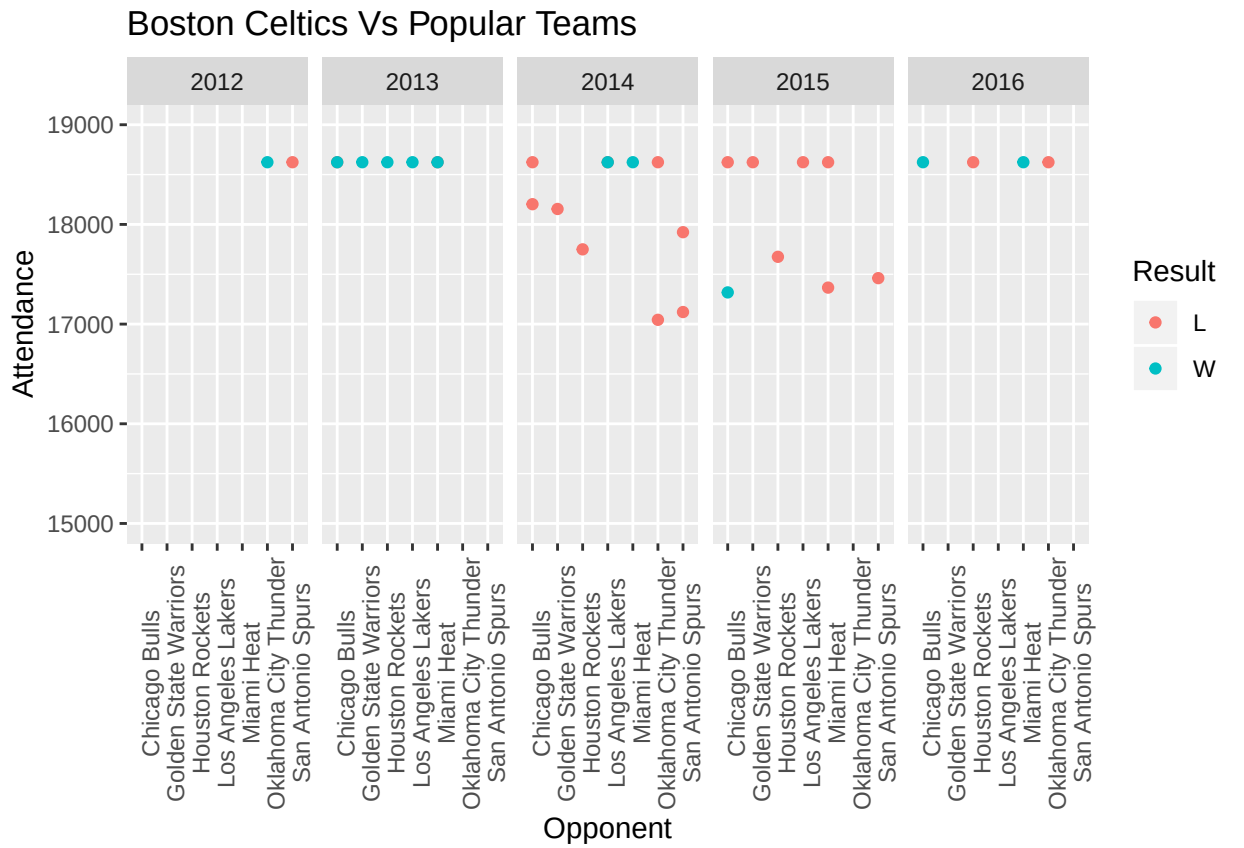


The plot shows that the presence of a star player in the opponent team is accompanied with a consistent and slightly higher attendance compared to not having a star player in the opponent team.

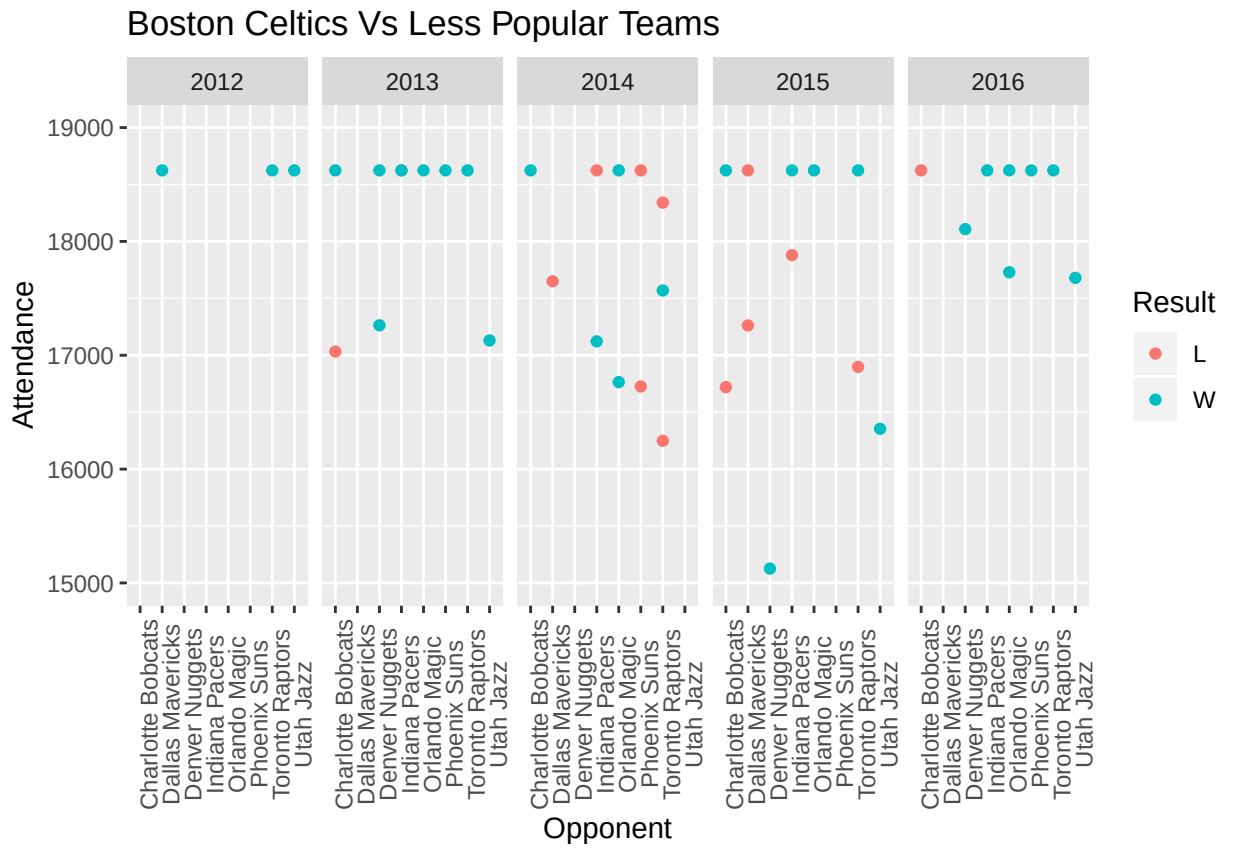
```
library(dplyr)
#Grouping the data by the opponent team and year
b_celtics <- BCeltics %>%
  mutate(year = substr(x= Date, start = 1, stop = 4)) %>%
  group_by(year, Opponent, Attendance, Result) %>%
  summarise()
b_celtics <- as.data.frame(b_celtics)

#Famous rivalries - Atlanta Hawks, Brooklyn Nets, Detroit Pistons, Los Angeles Lakers, New York Knicks,

#Boston Celtics Vs Rivals
a_h <- filter(b_celtics, b_celtics$Opponent %in% c("Atlanta Hawks", "Brooklyn Nets", "Detroit Pistons",
ggplot(a_h)+
  geom_point(aes(x = Opponent, y = Attendance, color = Result))+
  theme(axis.text.x = element_text(angle = 90))+
  facet_grid(.~year)+
  ylim(15000,19000)+
  ggtitle("Boston Celtics Vs Rivals")
```




```
#Boston Celtics Vs Popular Teams
b_h <- filter(b_celtics, b_celtics$Opponent %in% c("Golden State Warriors", "Chicago Bulls", "Los Angele
ggplot(b_h)+
  geom_point(aes(x = Opponent, y = Attendance, color = Result))+
  theme(axis.text.x = element_text(angle = 90))+
  facet_grid(.~year)+
  ylim(15000,19000)+
  ggtitle("Boston Celtics Vs Popular Teams")
```




```
#Boston Celtics Vs Less Popular Teams
c_h <- filter(b_celtics, b_celtics$Opponent %in% c("Toronto Raptors", "Denver Nuggets", "Charlotte Bobca
ggplot(c_h)+
  geom_point(aes(x = Opponent, y = Attendance, color = Result))+
  theme(axis.text.x = element_text(angle = 90))+
  facet_grid(.~year)+
  ylim(15000,19000)+
  ggtitle("Boston Celtics Vs Less Popular Teams")
```



We can see that the plot for Boston Celtics Vs Popular Teams has a consistent and slightly higher attendance (above 17000), comapred to Boston Celtics Vs Less Popular Teams. Though it is not a large variation, spectators seem to enjoy games between popular teams more. Existing rivalry between the teams does not seem to have much effect on spectator attendance. Notice that the game Celtics Vs Denver Nuggets in 2015 has an unusually low attendance. This may be because the Celtics lost the previous four games (as shown in the table below).

```
  library(knitr)
include_graphics("Capture.png")
```

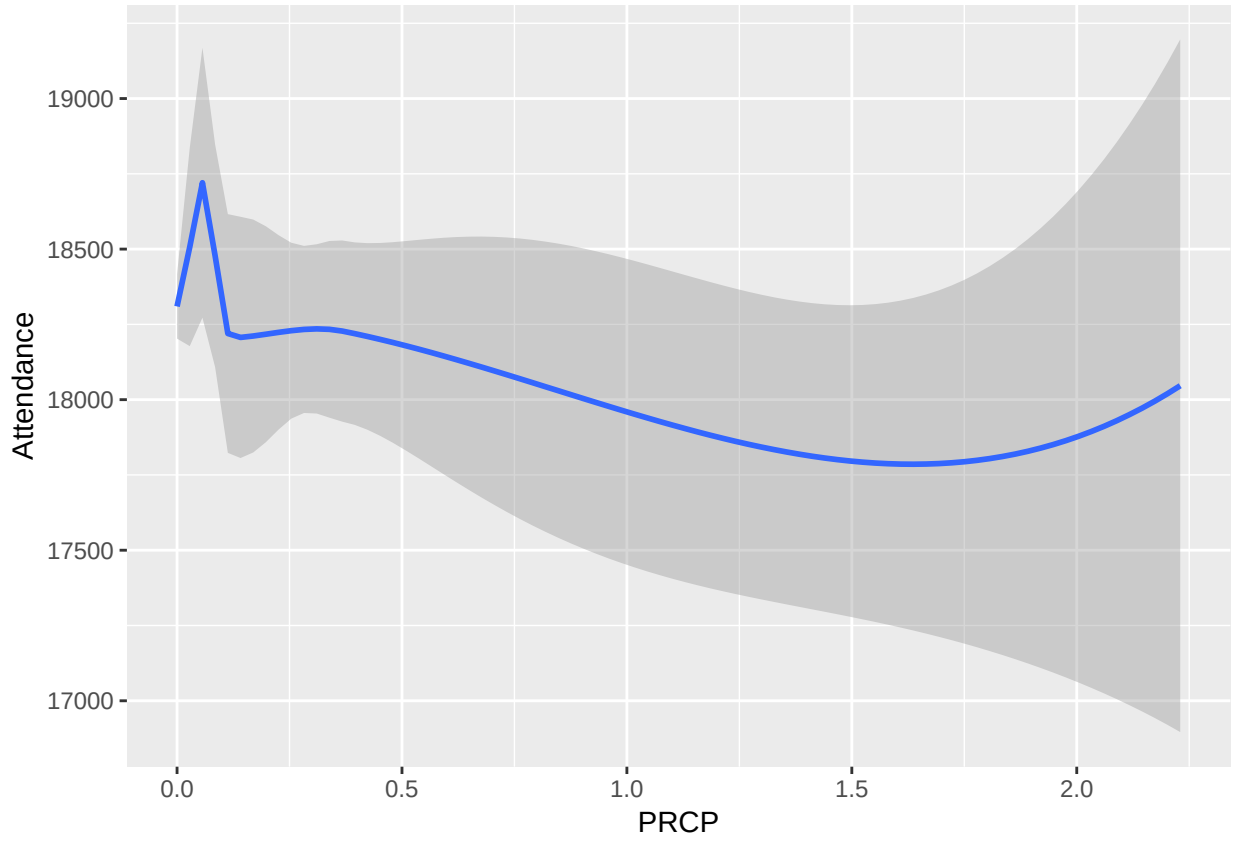
Date	Team	Attendance	Opponent	OSP	Result
2015-01-05	Boston Celtics	16720	Charlotte Bobcats	No	L
2015-01-12	Boston Celtics	16905	New Orleans Pelicans	Yes	W
2015-01-14	Boston Celtics	16067	Atlanta Hawks	No	L
2015-01-16	Boston Celtics	18624	Chicago Bulls	No	L
2015-01-30	Boston Celtics	17675	Houston Rockets	No	L
2015-02-01	Boston Celtics	17366	Miami Heat	No	L
2015-02-04	Boston Celtics	15126	Denver Nuggets	No	W

(C) Comparisons

We focus on the different effect of rain that brings to the attendance at Celtics and Red Sox. From the first graph, we can find the basketball attendance has more obvious trends that, with the increase in rain, the attendances tend to decrease. However, this phenomenon does not fit our consensus.

```
library(dplyr)
table_celtics<-Celtics
table_celtics$Attend.<-as.numeric(gsub(",","",table_celtics$Attend.))
colnames(table_celtics)[3]="Attendance"
join_rain_celtics<-inner_join(table_celtics,weather,by="Date")
p_c4<-ggplot(data = join_rain_celtics)+geom_smooth(mapping = aes(x=PRCP,y=Attendance))
p_c4
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
p_red4
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at -0.0088

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 0.0288

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 1.3804e-015

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
```

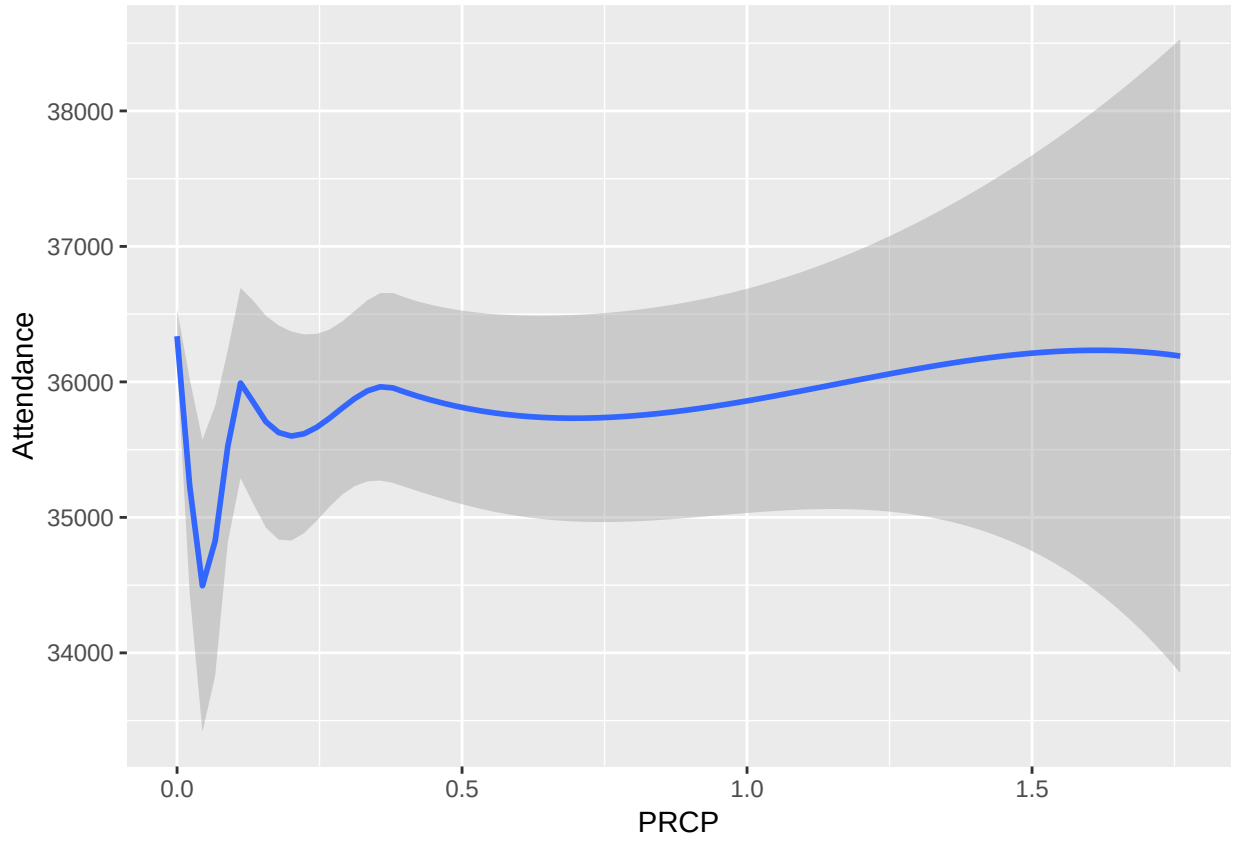
```
## parametric, : There are other near singularities as well. 0.0004
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used
## at -0.0088
```

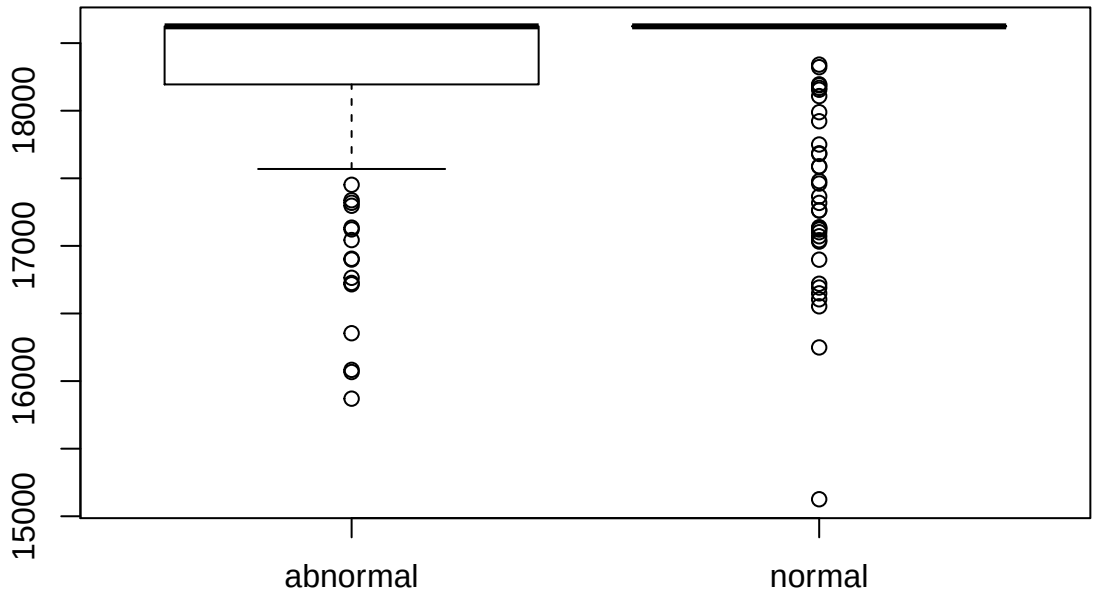
```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 0.0288
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal
## condition number 1.3804e-015
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other
## near singularities as well. 0.0004
```




```
##Secondly, we try to figure out the difference between basketball and baseball under different weather

table_type_c<-inner_join(weather,table_celtics,by="Date")
boxplot(data=table_type_c,Attendance~sum)
```



From the analysis above, we can find that the basketball attendances are less likely to be affected compared to baseball attendances.

‘‘‘