MA615 midterm project

Zhaobin Liu, Yifu Dong, Xiangliang Liu, Jinfei Xue

Oct 18 2018

Introduction:

This report will explore the relationship between the basketball & baseball attendace and weather condition. Specifically, we are exploring the influence of wind speed, average temperature, average precipitaion and weather types(Heavy fog, Thunder, Smoke or haze, Blowing and drifting snow) on the audience's attendance. The data are mainly exported from the The whole will include data scraping, data cleaning, data visualization and shiny app.

Data revriving:

Boston Red Socks attendance is retrived from: https://www.baseball-reference.com/teams/BOS/2017-schedule-scores.shtml (https://www.baseball-reference.com/teams/BOS/2017-schedule-scores.shtml) Boston Celtics attendance is retrived from: http://www.espn.com/nba/game?gameId=400828055 (http://www.espn.com/nba/game?gameId=400828055) Weather condition information is retrived from: https://www.ncdc.noaa.gov/cdo-web/search (https://www.ncdc.noaa.gov/cdo-web/search)

Package:

```
#install.packages("rvest")
#install.packages("XML")
#install.packages("RCurl")
#install.packages("stringr")
library(rvest)
```

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

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────── tidyverse 1.2.1 ──
```

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

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

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

Data scraping and data cleaning:

```r
# Baseball Data

# Scrape the data
url1 <- "https://www.baseball-reference.com/teams/BOS/"
url2 <- "-schedule-scores.shtml"
years <- c(2012:2017)
urls <- str_c(url1, years, url2, sep = "")
filenames <- str_c("baseball", years, sep = "")

for (i in 1:length(urls)) {
  read_url <- read_html(urls[i])
  file = read_url %>%
         html_table(fill=TRUE)%>%
         .[[1]]
  suppressMessages(
    assign(filenames[i], file)
  )

  colnames(file)[1] <-"YYYY"
  colnames(file)[5] <-"home"
  file = file[!str_detect(file$YYYY,"Gm#"),]
  file[,1] = years[i]

  if(i == 1){
    baseball <- file
  }
  else{
    baseball <- rbind.data.frame(baseball, file)
  }

}

# Clean the data
baseball = baseball[!str_detect(baseball$home,"@"),]
baseball$Date = str_c(baseball$Date, baseball$YYYY, sep = ",")
baseball$Date = str_replace(baseball$Date, " \\(.*\\)", "")
baseball$Date = as.Date(baseball$Date,format="%a, %b %d,%Y")
baseball$Attendance = gsub(",","",baseball$Attendance)
baseball$Attendance = as.numeric(as.character(baseball$Attendance))


# Export the data as csv
#getwd()
#setwd("D:/2018_Semester_1/MA615 Data Science in R/B1-Lecture/Assignment/Midterm_proj
ect/data")
#write.csv(baseball, file = "baseball.csv", row.names = F, quote = F)
```

```r
# Basketball Game Date
# Scrape tables on the Internet
```

```
years <- 2012:2018
urls <- paste0("http://www.espn.com/nba/team/schedule/_/name/bos/season/", years, "/s
easontype/2")

get_table <- function(url) {
  url %>%
    read_html() %>%
    html_nodes(xpath = '/html/body/div[1]/div/div/div/div/div[5]/div[3]/div[2]/div[1]
/div[1]/article/div/section/div[2]/section/section/table/tbody/tr/td/div/div/div[2]/t
able/tbody/tr/td/table') %>%
    html_table(fill = TRUE)
}

results <- sapply(urls, get_table)

# Delete the first two rows
nrow <- rep(NULL, length(results))
for (i in 1:length(results)){
  results[[i]] <- results[[i]][-(1:2), 1:3]
  rownames(results[[i]]) <- 1:nrow(results[[i]])
  nrow[i] <- dim(results[[i]])[1]
}

# Combine the six dataframes
for (i in 1:length(results)){
  if(i == 1){
   data_bask <- results[[1]][, 1:3]
  }
  else{
  data <- results[[i]][, 1:3]
  data_bask <- rbind.data.frame(data_bask, data)
  }
}
names(data_bask)<- c("DATE", "OPPONENT","RESULT")


# First eliminate games in 2018 and 2011
data_2018 <- nrow(data_bask)- which(results[[7]][41:nrow[7],] == "Wed, Apr 11")+1
data_bask <- data_bask[-(data_2018:nrow(data_bask)),]
data_2011 <- which(results[[1]][1:nrow[1],] == "Fri, Dec 30")
data_bask <- data_bask[-(1:data_2011),]

# Add column "YYYY" in dataframe

end_2012 <- which(results[[1]][5:nrow[1],] == "Thu, Apr 26")+
  which(results[[2]][1:nrow[2],] == "Sun, Dec 30")

end_2013 <- which(results[[2]][31:nrow[2],] == "Wed, Apr 17")+
  which(results[[3]][1:nrow[3],] == "Tue, Dec 31")
```

```
end_2014 <- which(results[[3]][32:nrow[3],] == "Wed, Apr 16")+
  which(results[[4]][1:nrow[4],] == "Wed, Dec 31")

end_2015 <- which(results[[4]][30:nrow[4],] == "Wed, Apr 15")+
  which(results[[5]][1:nrow[5],] == "Wed, Dec 30")

end_2016 <- which(results[[5]][33:nrow[5],] == "Wed, Apr 13")+
  which(results[[6]][1:nrow[6],] == "Fri, Dec 30")

end_2017 <- which(results[[6]][35:nrow[6],] == "Wed, Apr 12")+
  which(results[[7]][1:nrow[7],] == "Sun, Dec 31")

YYYY <- rep(2012:2017,c(end_2012, end_2013, end_2014, end_2015, end_2016, end_2017))
data_bask <- cbind.data.frame(YYYY, data_bask)


# Delete the row of canceled and postponed games
n = grep("Canceled", data_bask$RESULT)
data_bask <- data_bask[-n,]
m = grep("Postponed", data_bask$RESULT)
data_bask <- data_bask[-m,]


# Transform format of "DATE"
data_bask$DATE = str_c(data_bask$DATE, data_bask$YYYY, sep = ",")
data_bask$DATE = str_replace(data_bask$DATE, " \\(.*\\)", "")
data_bask$DATE = as.Date(data_bask$DATE,format="%a, %b %d,%Y")
```

Joining weather data and game attendance data

We can directly get the xlsx file throught the website. Then we can join the weather data and attendance data by data:

```
library(readxl)
weather  <- read_excel("weather.xlsx")
weather <- na.omit(weather)

#finalbaseball <- inner_join(baseball,weather,by="DATE")
#finalbasket <- inner_join(data_bask,weather,by="DATE")
```

Visualization(EDA) and disscussion:

Attendance and average wind speed(Xiangliang Liu)

```
#read in raw data:
finalbasket <- read_csv("finalbasket.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   STATION = col_character(),
##   NAME = col_character(),
##   DATE = col_character(),
##   `AWND(avg wind speed)` = col_double(),
##   `PRCP(precipitation)` = col_double(),
##   `SNOW(snowfall)` = col_double(),
##   `TAVG(temprature avg)` = col_double(),
##   WSF2 = col_double(),
##   Date = col_character(),
##   OPPONENT = col_character(),
##   RESULT = col_character(),
##   date = col_character()
## )
```

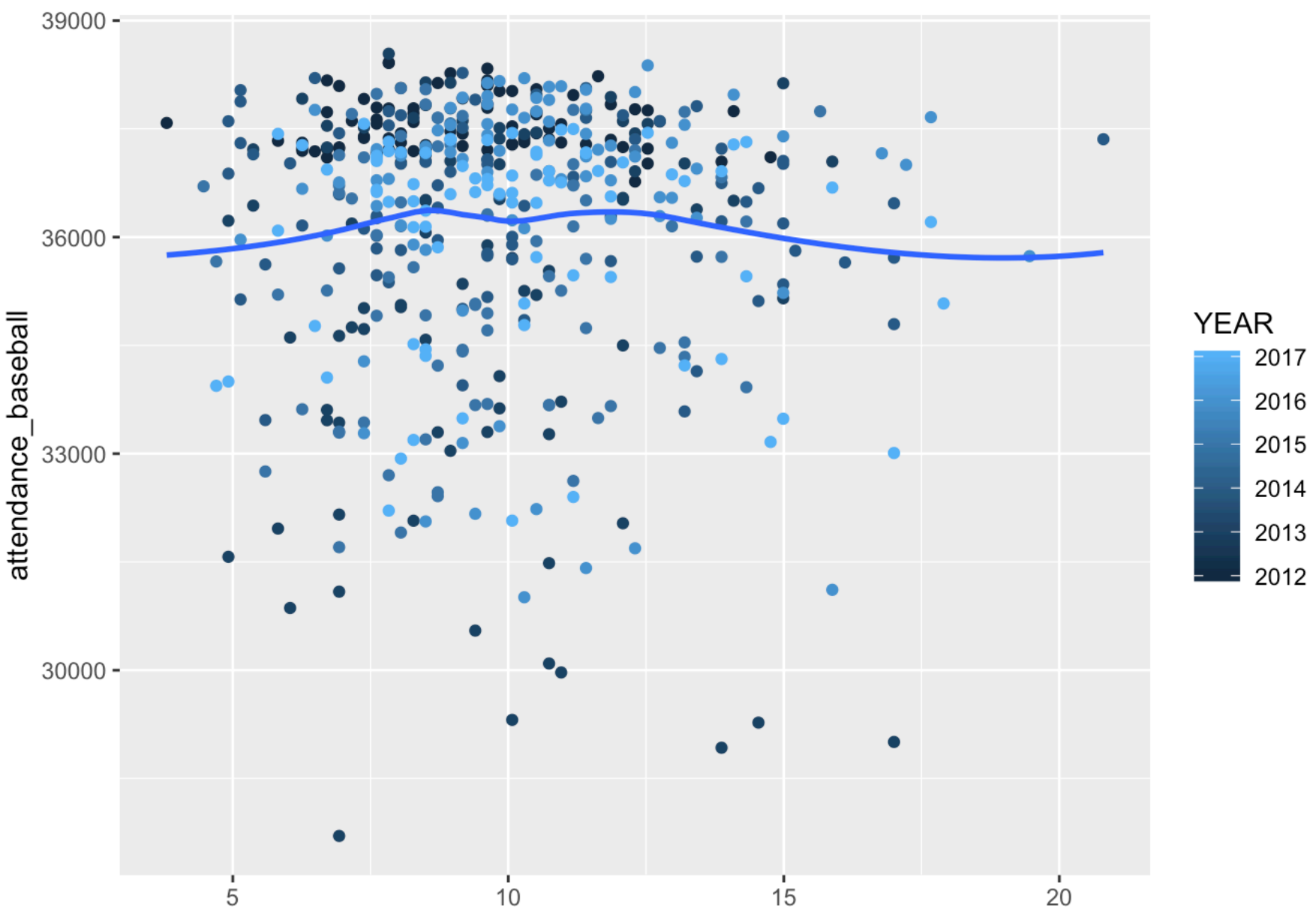
```
## See spec(...) for full column specifications.
```

```
basketball = finalbasket[,c(1:8, 29, 33)] #subset basketball data.
finalbaseball <- read_csv("~/Downloads/finalbaseball.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   STATION = col_character(),
##   NAME = col_character(),
##   DATE = col_character(),
##   `AWND(avg wind speed)` = col_double(),
##   `PRCP(precipitation)` = col_double(),
##   `TAVG(temprature avg)` = col_double(),
##   WSF2 = col_double(),
##   Date = col_character(),
##   Var.28 = col_character(),
##   Tm = col_character(),
##   home = col_character(),
##   Opp = col_character(),
##   `W/L` = col_character(),
##   `W-L` = col_character(),
##   GB = col_character(),
##   Win = col_character(),
##   Loss = col_character(),
##   Save = col_character(),
##   Time = col_time(format = ""),
##   `D/N` = col_character()
##   # ... with 2 more columns
## )
## See spec(...) for full column specifications.
```

```
baseball = finalbaseball[,c(1:8, 27, 45)] #subset baseball data.
names(baseball) = c("NUMBER","STATION","NAME","DATE","AWND","PRCP","SNOW","TAVG_baseb
all","YEAR","attendance_baseball")
names(basketball) = c("NUMBER","STATION","NAME","DATE","AWND","PRCP","SNOW","TAVG_bas
ketball","YEAR","attendance_basketball")
#ggplot of Attendance and average wind speed on baseball
ggplot(data = baseball, mapping = aes(x = AWND, y =attendance_baseball, color=YEAR))
+ geom_point() + geom_smooth(mapping = aes(x =AWND, y = attendance_baseball), se =FAL
SE)
```

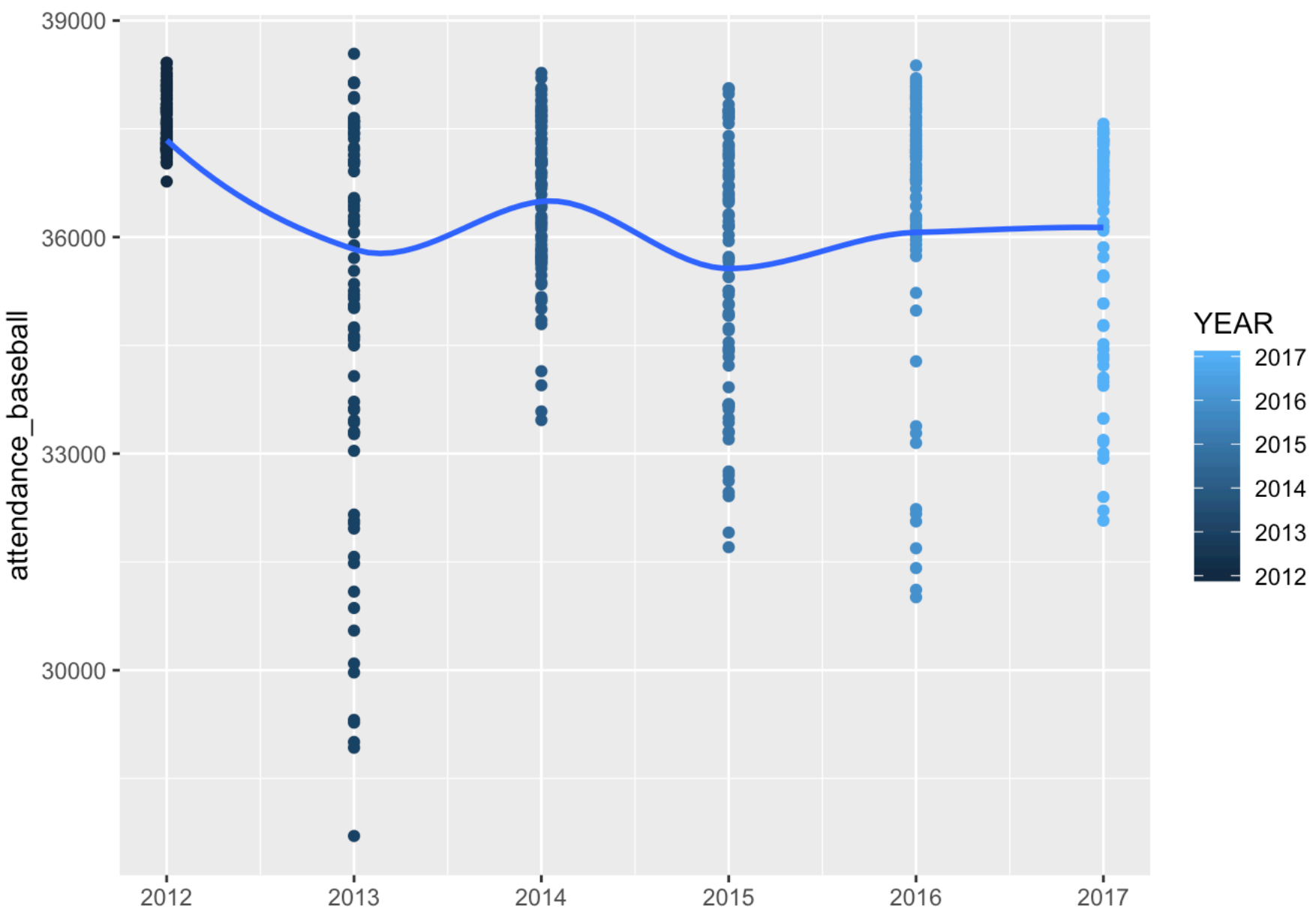
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



AWND

```
#ggplot of average baseball attendance from 2012 to 2017
ggplot(data = baseball, mapping = aes(x = YEAR, y =attendance_baseball, color=YEAR))
+ geom_point() + geom_smooth(mapping = aes(x =YEAR, y = attendance_baseball), se =FAL
SE)
```

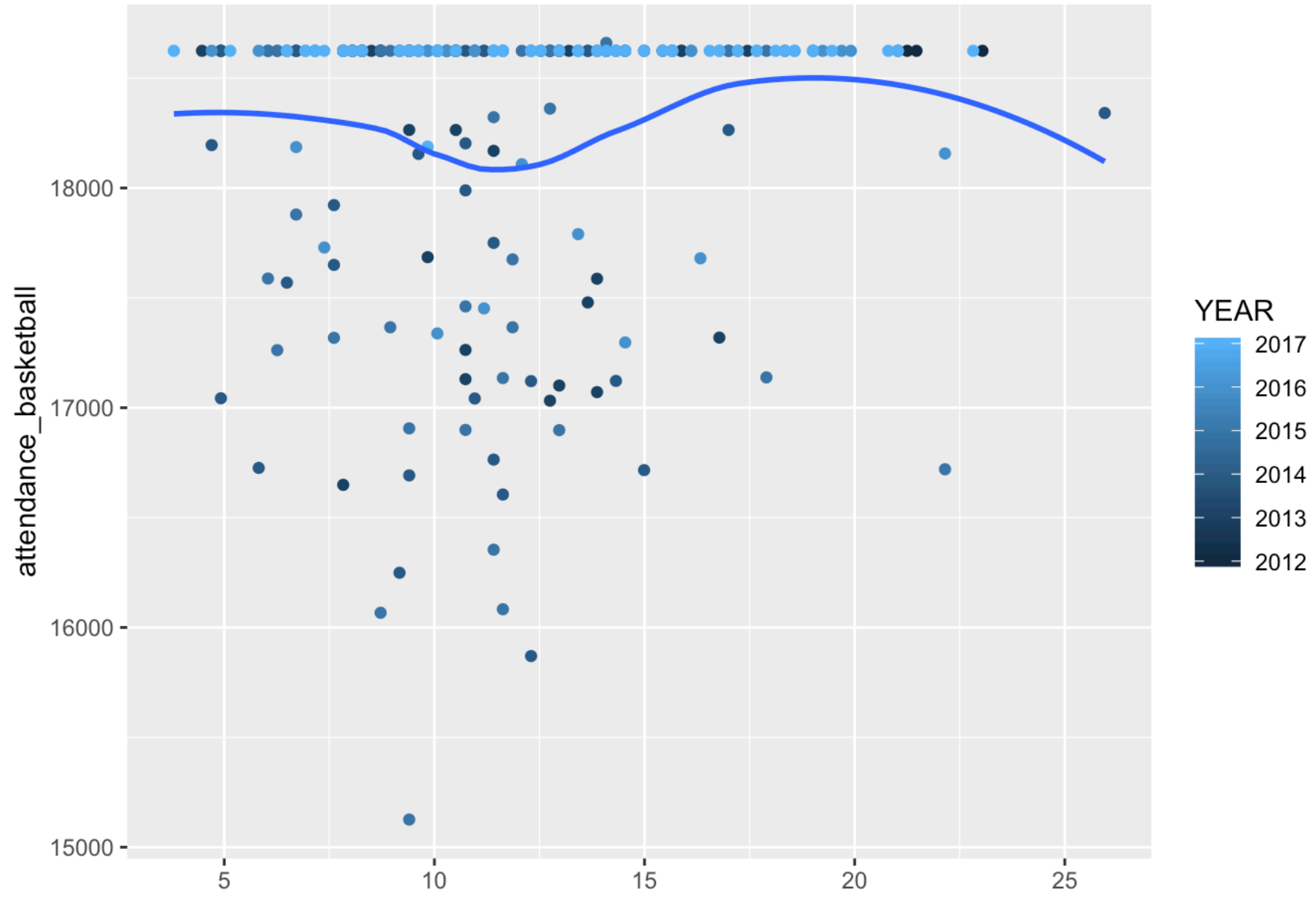
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



YEAR

```
#ggplot of attendance and average wind speed on basketball
ggplot(data = basketball, mapping = aes(x = AWND, y = attendance_basketball, color=YE
AR)) + geom_point() + geom_smooth(mapping = aes(x = AWND, y = attendance_basketball )
,se = FALSE)
```

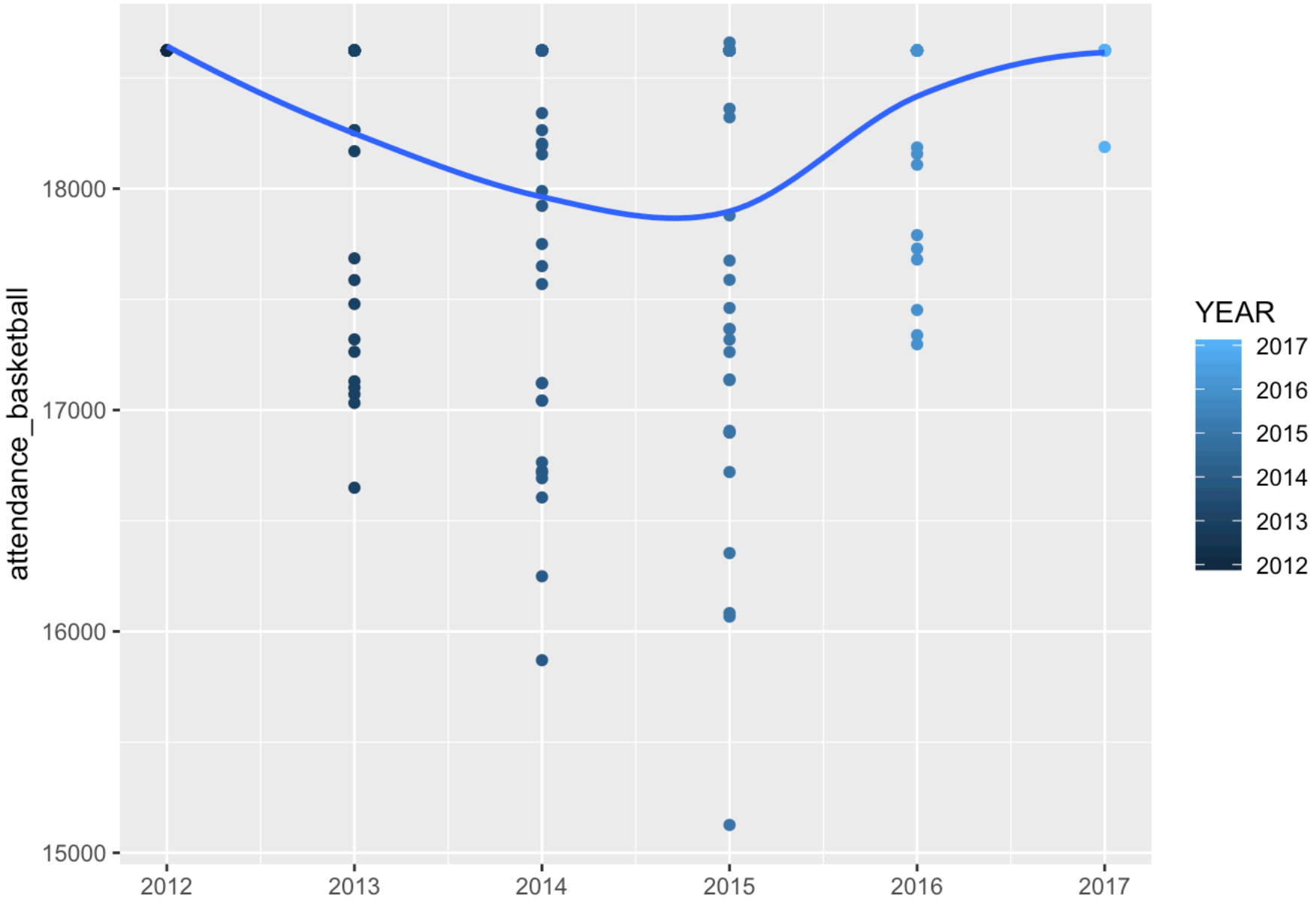
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



AWND

```
#ggplot of average basketball attendance from 2012 to 2017
ggplot(data = basketball, mapping = aes(x = YEAR, y =attendance_basketball, color=YEA
R)) + geom_point() + geom_smooth(mapping = aes(x =YEAR, y = attendance_basketball), s
e =FALSE)
```

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



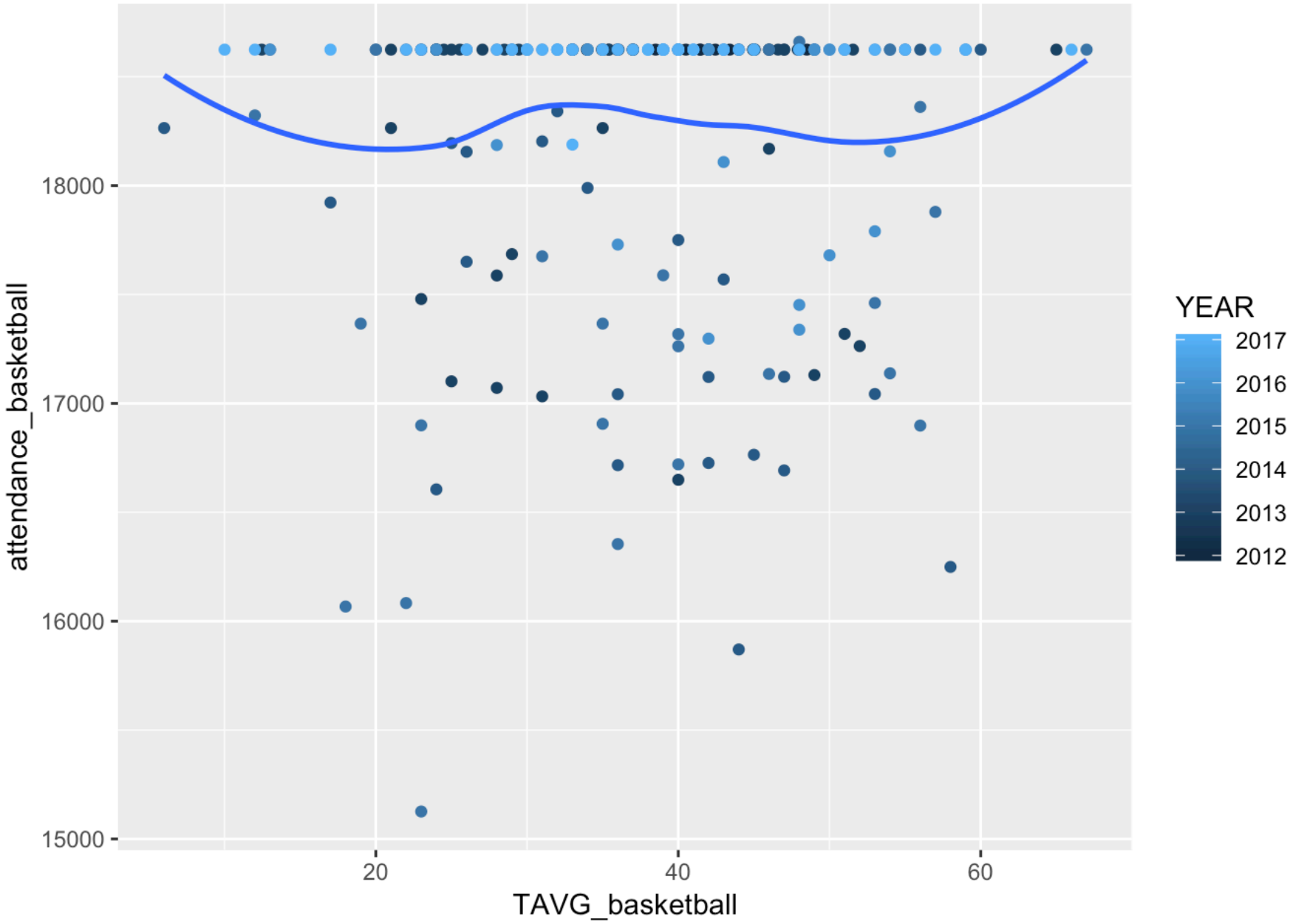
YEAR

#From the ggplot above, 2015 seems have lowest attendance rate on both basketball and baseball.
#on the attendance VS average wind speed plots, there's no clear linear relationship between attendance and average wind speed. But we can see that as average wind speed increasing, the attendance increased first and then decreased. This is reasonable. Since baseball is an outdoor sports, people are more willing to enjoy the game when there is breeze. But people tend to quit attending the sports when there are very strong wind outside.By checking the ggplot of basketball attedndance and wind speed, we found there's no relationship between those two terms. This is mainly because basketball games are held indoor. Audience are less influenced by wind speed.

Attendance and average temperature(Zhaobin Liu)

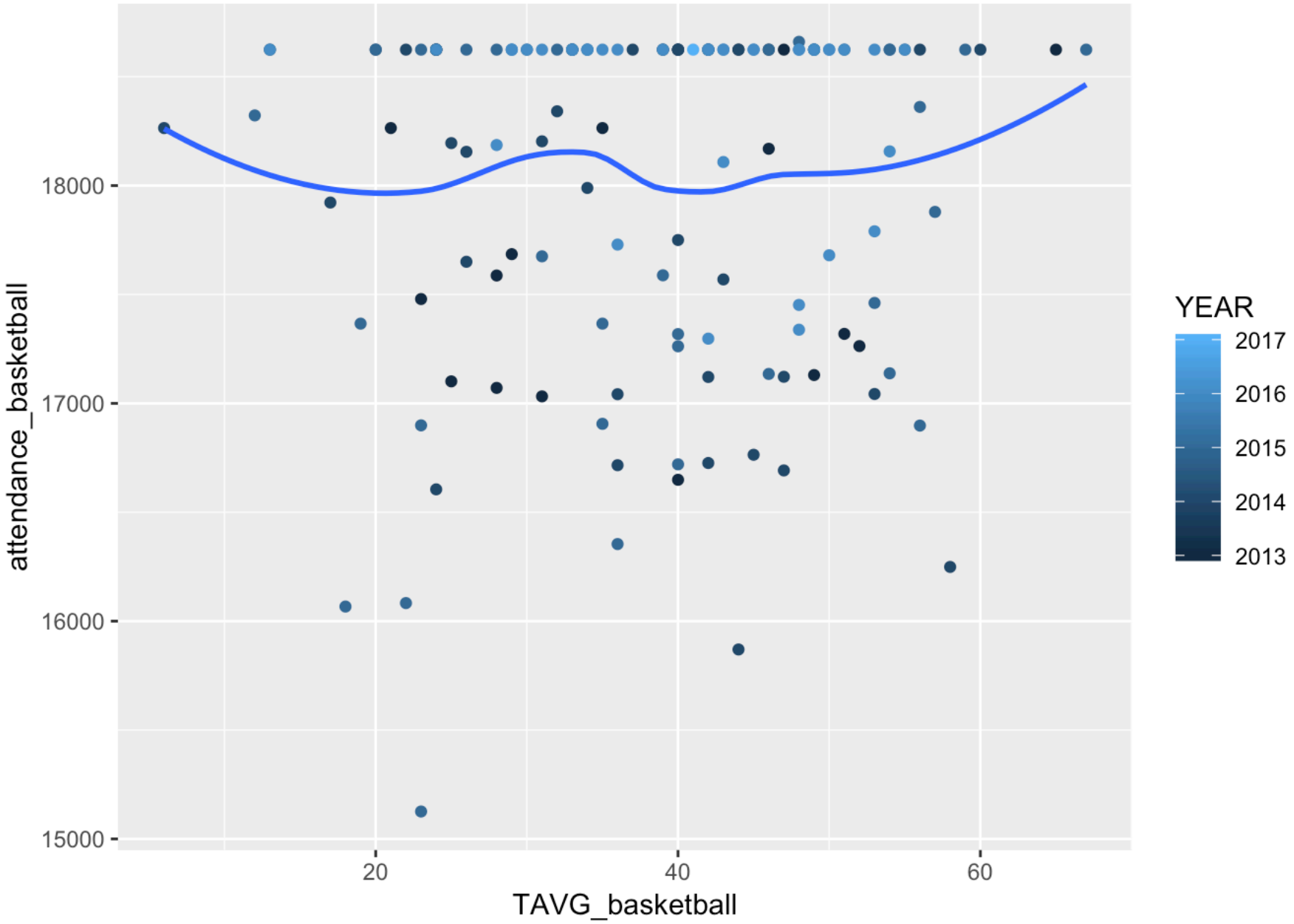
```
ggplot(data = basketball, mapping = aes(x = TAVG_basketball, y = attendance_basketbal
l, color = YEAR)) + geom_point() +geom_smooth(mapping =aes(x = TAVG_basketball, y = a
ttendance_basketball ), se = FALSE)
```

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



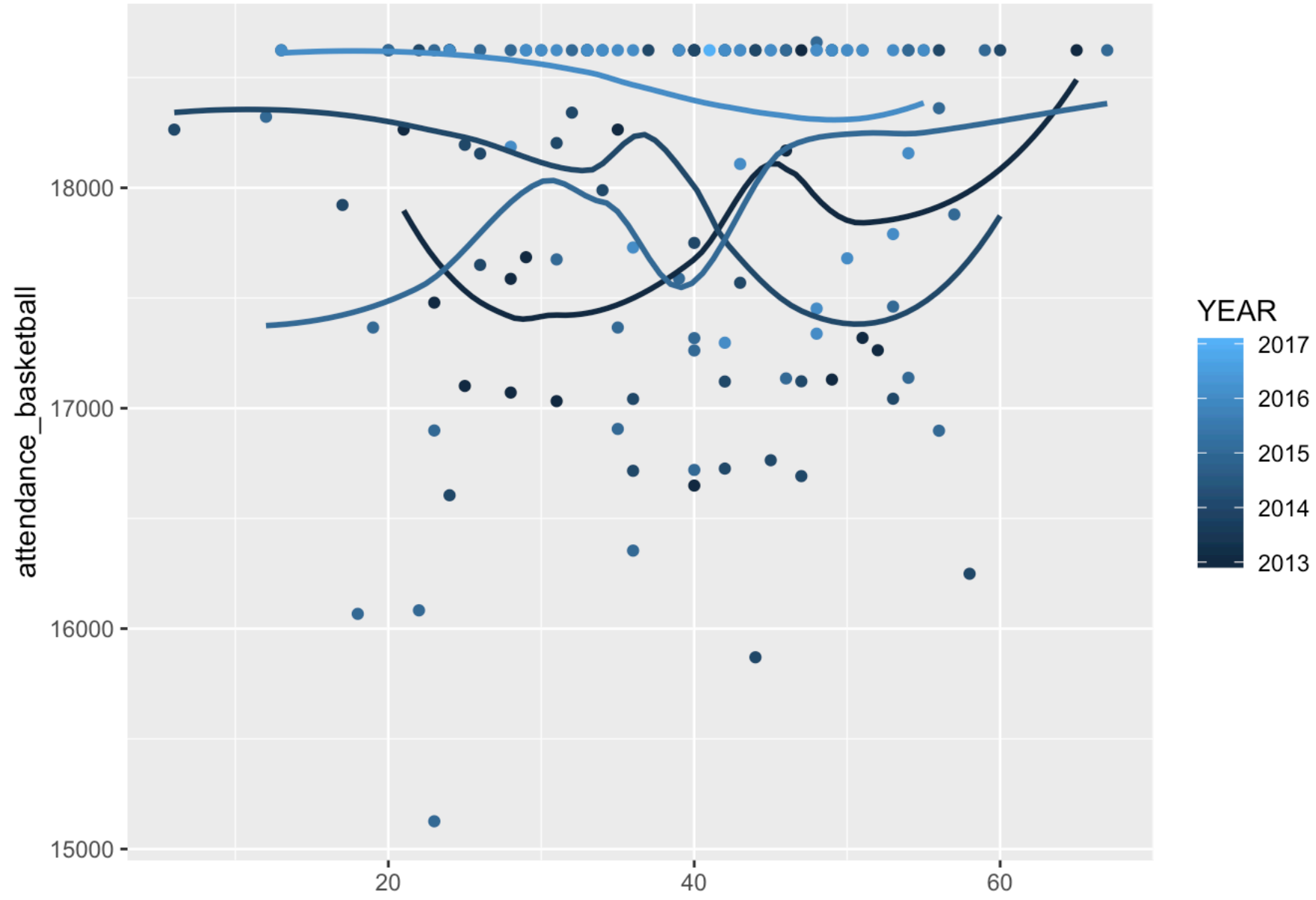

```
#basketball in the mid three year
#select the data from 2013 to 2016
finalbasket_three = basketball[41:179,]
ggplot(finalbasket_three, aes(x=TAVG_basketball, y=attendance_basketball, color=YEAR)
) + geom_point()+
  geom_smooth(mapping = aes(x=TAVG_basketball, y=attendance_basketball),se=FALSE)
```

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



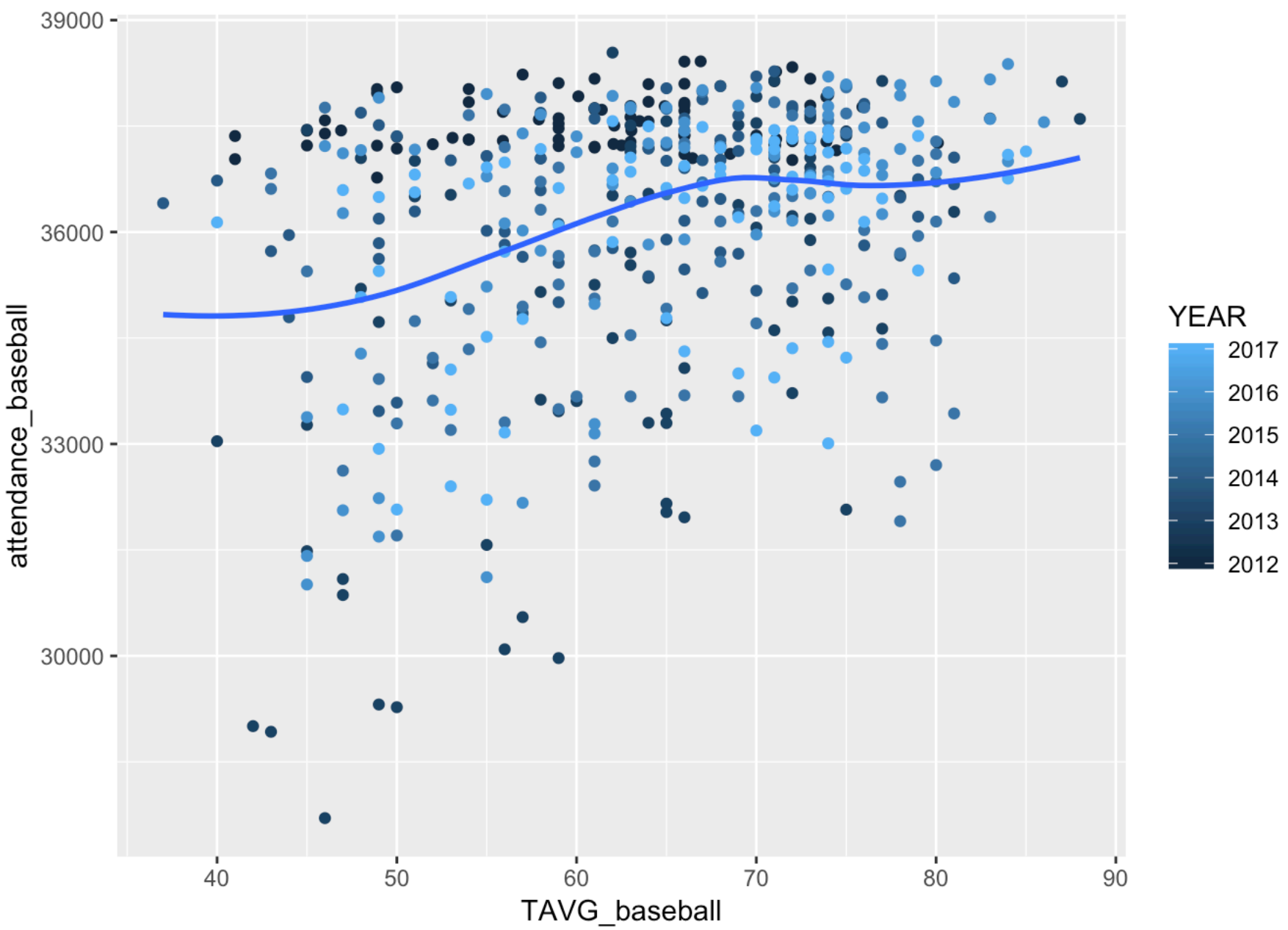

```
#Average temperature vs basketball attendance for 2012 to 2017
ggplot(finalbasket_three, aes(x=TAVG_basketball, y=attendance_basketball, color=YEAR)
) + geom_point()+
  geom_smooth(mapping = aes(x=TAVG_basketball, y=attendance_basketball, group=YEAR),s
e=FALSE)
```

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



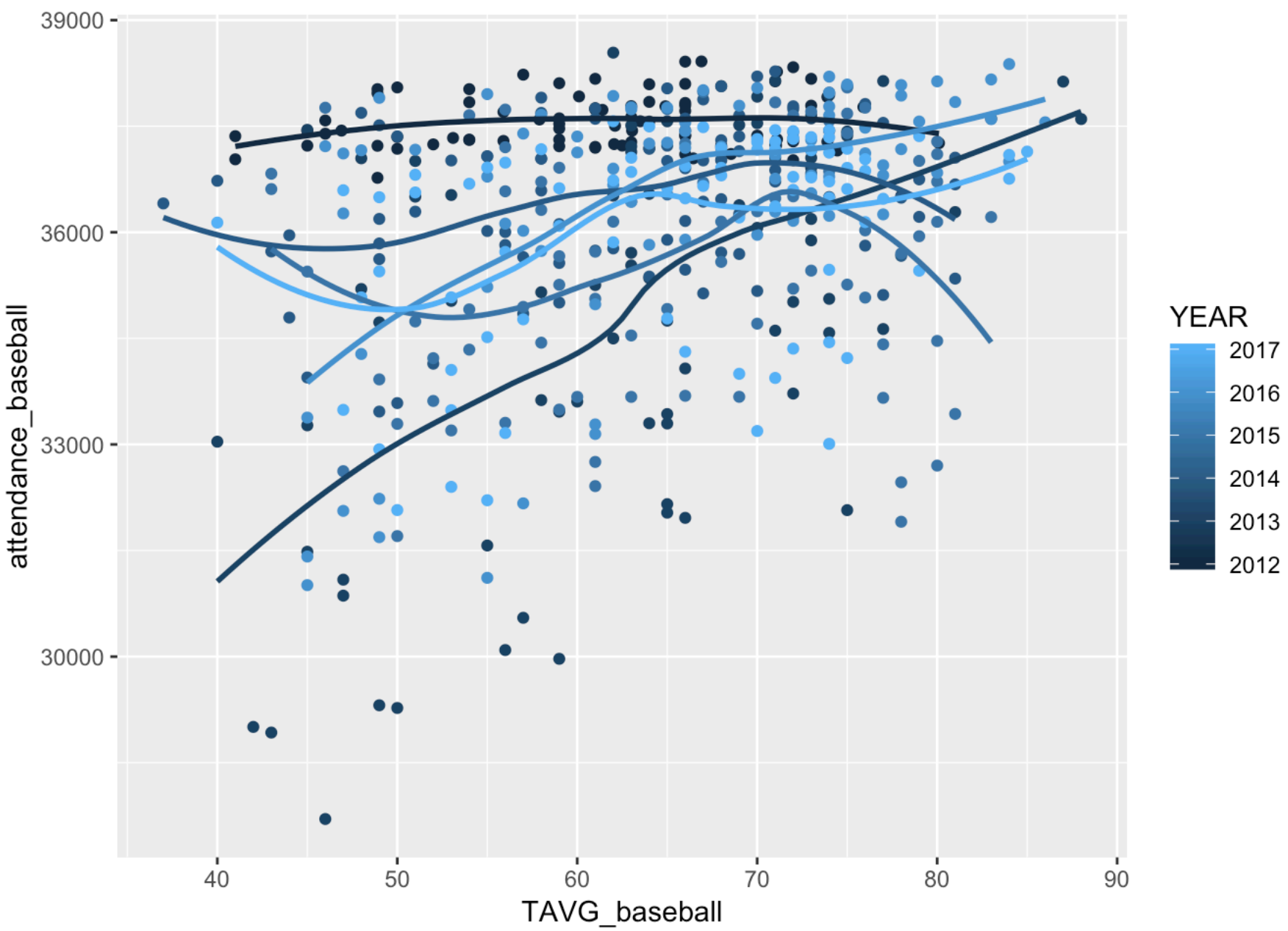

```
#baseball
ggplot(baseball, aes(x=TAVG_baseball, y=attendance_baseball, color=YEAR)) + geom_poin
t()+
  geom_smooth(mapping = aes(x=TAVG_baseball, y=attendance_baseball),se=FALSE)
```

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




```
ggplot(baseball, aes(x=TAVG_baseball, y=attendance_baseball, color=YEAR)) + geom_poin
t()+
  geom_smooth(mapping = aes(x=TAVG_baseball, y=attendance_baseball, group=YEAR),se=FA
LSE)
```

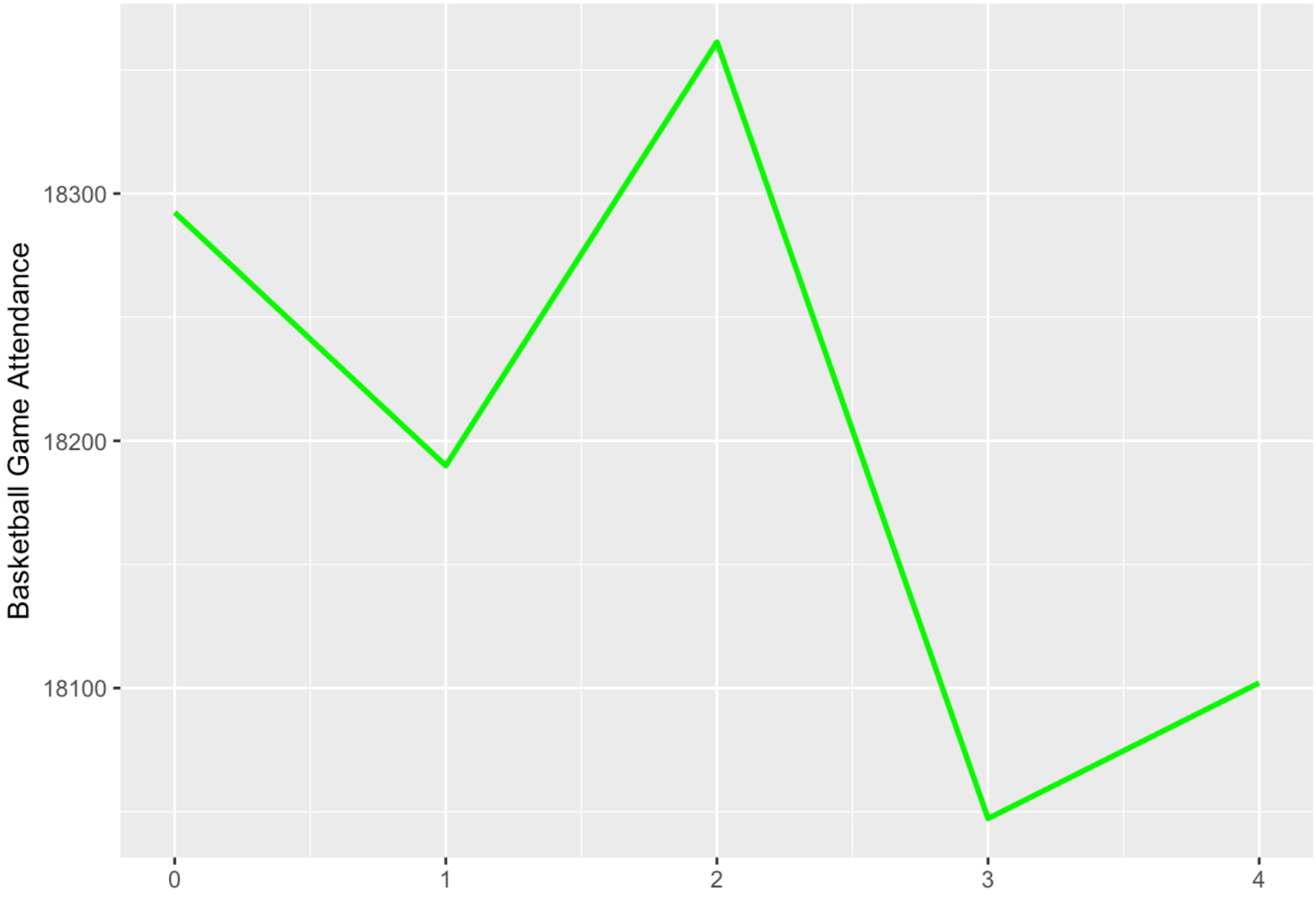
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



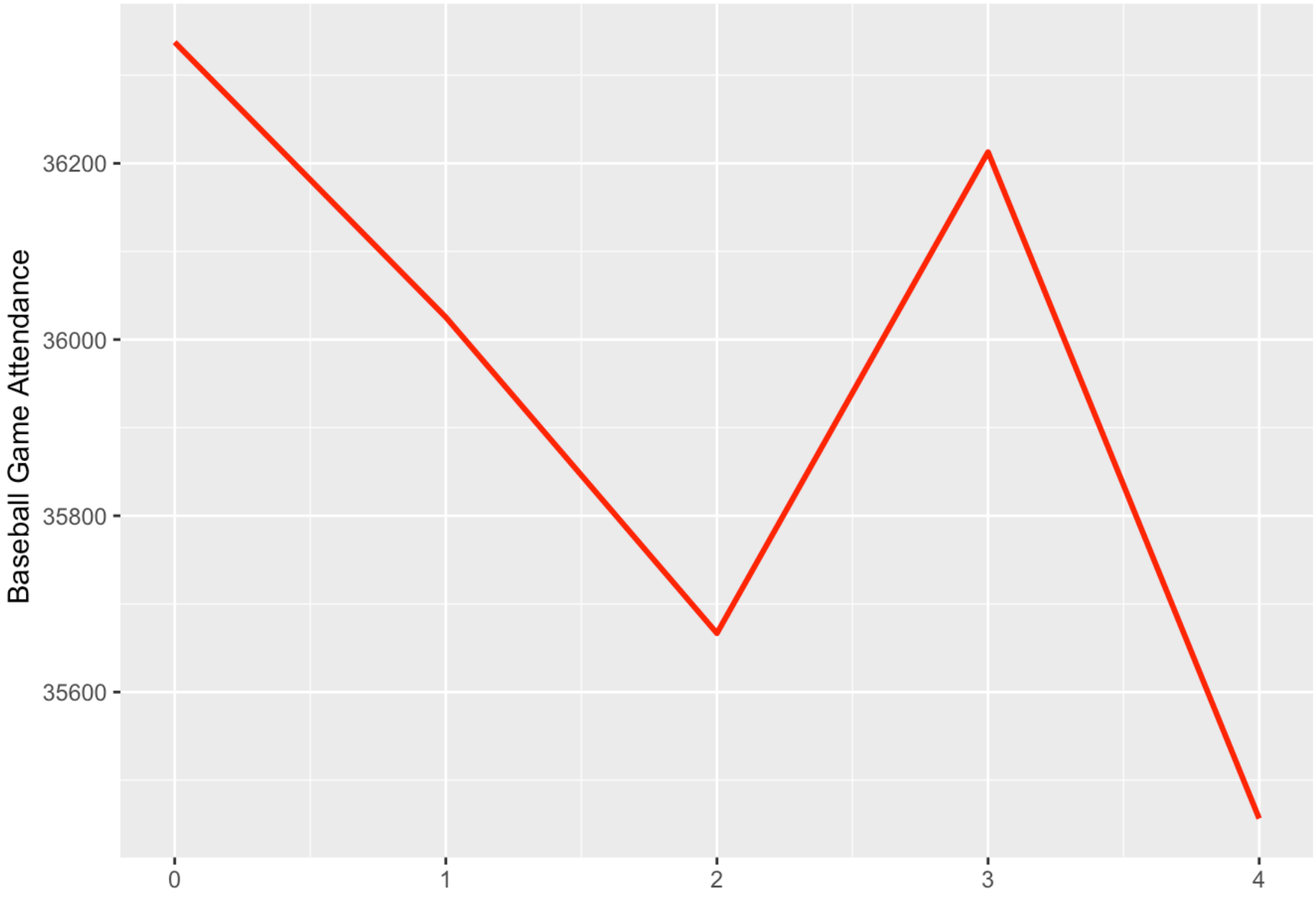
First we will analyze the relationship between attendance and temperature in baseball game. By graph, we can see the attendance trend of the interval between 40 and 70 are going up showing that there might be some positive relationship between temperature and attendance. Later on, the weather does not affect the attendance a lot. At last, we graph the trend of all six years. We can see when the temperature is low, the attendance is not very high. While the temperature is going up, the attendance will be increased until some point that either the increasing rate will be slower or the attendance will start to decrease. For basketball, we are unable to see the clear trend between the temperature and attendance using all years long data. Then we decided to cut the data of the first year and the last year(2012 and 2017) since there exist too many full attendance which is 18624. After we graph it, it seems like there is still not any relationship between them. At last, we graph each year's trend and compare. Still, there is still no clear relationship. Thus, we conclude that temperature will not have a great effect of attendance rate.

Attendance and precipitation data(Jinfei Xue)

```
#Basketball
finalbasket <- read.csv("finalbasket.csv")
bask.prcp <- finalbasket %>%
  select(DATE, PRCP.precipitation., ATTENDANCE)

# Add a catogorical variable of precipitation
mean.prcp <- mean(bask.prcp$PRCP.precipitation.)
sd.prcp <- sd(bask.prcp$PRCP.precipitation.)
# 0 stands for no precipitation; from 1 to 4, the precipitation becomes more and more
.
bask.prcp$prcp <- bask.prcp$PRCP.precipitation.
bask.prcp$prcp[0 < bask.prcp$PRCP.precipitation. & bask.prcp$PRCP.precipitation. < me
an.prcp] = 1
bask.prcp$prcp[mean.prcp <= bask.prcp$PRCP.precipitation. & bask.prcp$PRCP.precipitat
ion. < (mean.prcp + sd.prcp)] = 2
bask.prcp$prcp[(mean.prcp + sd.prcp) <= bask.prcp$PRCP.precipitation. & bask.prcp$PRC
P.precipitation. < (mean.prcp + 2*sd.prcp)] = 3
bask.prcp$prcp[bask.prcp$PRCP.precipitation. > (mean.prcp + 2*sd.prcp)] = 4

# Compute the mean attendance in each precipitation group
avg.bask <- bask.prcp %>%
  group_by(prcp) %>%
  summarise(avg.bask = mean(ATTENDANCE))

# Make ggplot for the relationship between degree of precipitation and basketball gam
e attendance
ggplot(data = avg.bask) +
  geom_line(mapping = aes(x=prcp, y=avg.bask), color='green', size = 1) +
  xlab("Degree of Precipitation") + ylab("Basketball Game Attendance")
```


Basketball Game Attendance
18300
18200
18100
0
1
2
3
4
Degree of Precipitation

```r
#Baseball
base <- read.csv("~/Downloads/finalbaseball.csv")
base.prcp <- base %>%
  select(DATE, PRCP.precipitation., Attendance)
mean.prcp <- mean(base.prcp$PRCP.precipitation.)
sd.prcp <- sd(base.prcp$PRCP.precipitation.)

# Add a catogorical variable of precipitation
base.prcp$prcp <- base.prcp$PRCP.precipitation.
base.prcp$prcp[0 < base.prcp$PRCP.precipitation. & base.prcp$PRCP.precipitation. < me
an.prcp] = 1
base.prcp$prcp[mean.prcp <= base.prcp$PRCP.precipitation. & base.prcp$PRCP.precipitat
ion. < (mean.prcp + sd.prcp)] = 2
base.prcp$prcp[(mean.prcp + sd.prcp) <= base.prcp$PRCP.precipitation. & base.prcp$PRC
P.precipitation. < (mean.prcp + 2*sd.prcp)] = 3
base.prcp$prcp[base.prcp$PRCP.precipitation. > (mean.prcp + 2*sd.prcp)] = 4

# Compute the mean attendance in each precipitation group
avg.base <- base.prcp %>%
  group_by(prcp) %>%
  summarise(avg.base = mean(Attendance))

# Make ggplot for the relationship between degree of precipitation and baseball game
attendance
ggplot(data = avg.base) +
geom_line(mapping = aes(x=prcp, y=avg.base), color='red', size = 1) +
xlab("Degree of Precipitation") + ylab(" Baseball Game Attendance")
```



Attendance and weather type (Yifu Dong)

```
#In our data, we divided weather types into several different catagories. For example
, snowfall, thunder, fog, smoke, haze, glaze, heavy fog and so on. Depend on the char
acteristic of Boston's weather, we decided to choose some typical weather types of Bo
ston:
#WT02 : Heavy fog or heaving freezing fog, ice
#WT03 : Thunder
#WT08 : Smoke or haze
#WT09 : Blowing and drifting snow

require(gridExtra)
```

```
## Loading required package: gridExtra
```

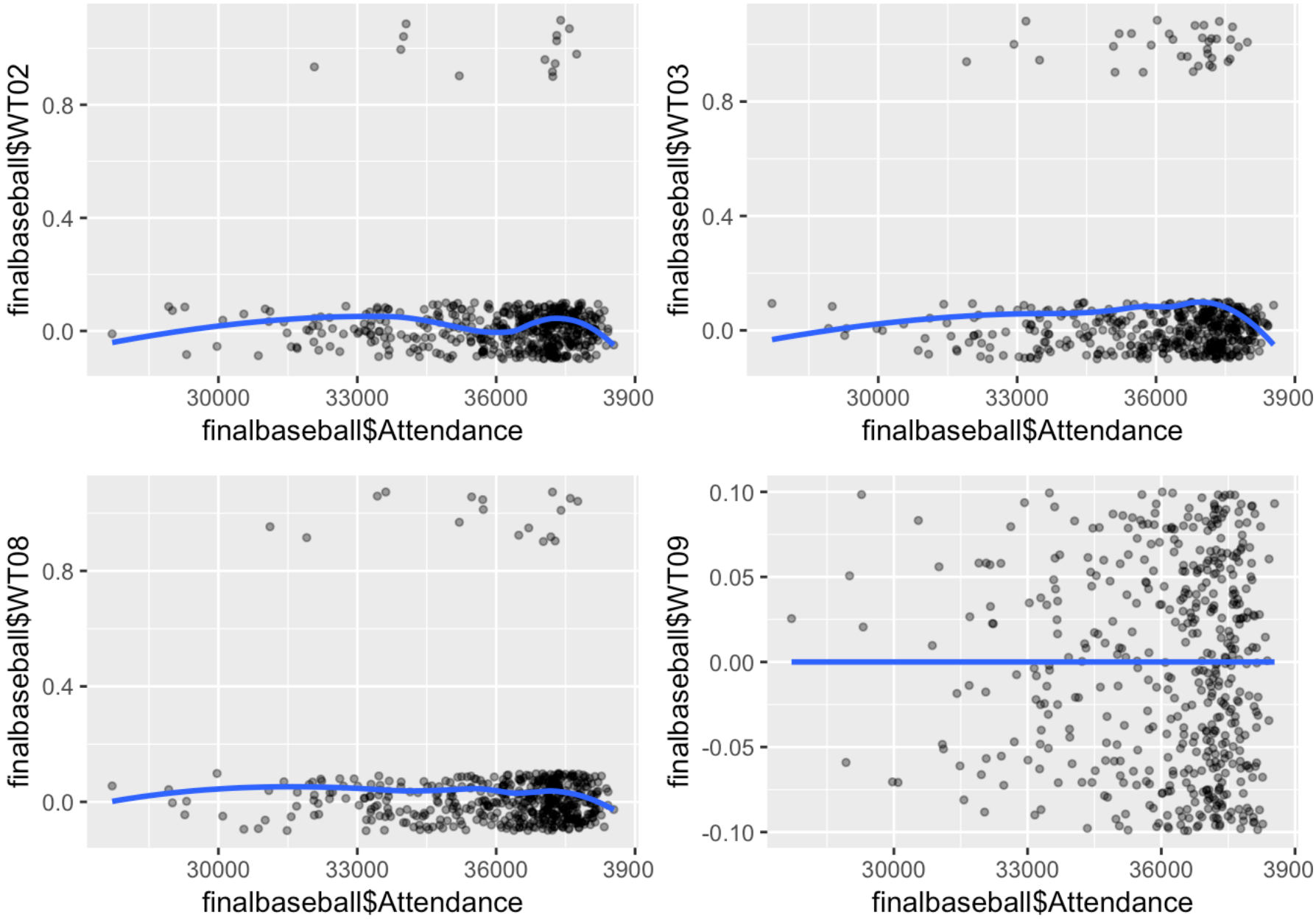
```
##
## Attaching package: 'gridExtra'
```

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

```
p1 <- ggplot(data = finalbaseball,mapping = aes(y=finalbaseball$WT02,x=finalbaseball$
Attendance))+
  geom_jitter(height=0.1,size=1, alpha=0.4)+
  geom_smooth(se = FALSE)
p2 <- ggplot(data = finalbaseball,mapping = aes(y=finalbaseball$WT03,x=finalbaseball$
Attendance))+
  geom_jitter(height=0.1,size=1, alpha=0.4)+
  geom_smooth(se = FALSE)
p3 <- ggplot(data = finalbaseball,mapping = aes(y=finalbaseball$WT08,x=finalbaseball$
Attendance))+
  geom_jitter(height=0.1,size=1, alpha=0.4)+
  geom_smooth(se = FALSE)
p4 <- ggplot(data = finalbaseball,mapping = aes(y=finalbaseball$WT09,x=finalbaseball$
Attendance))+
  geom_jitter(height=0.1,size=1, alpha=0.4)+
  geom_smooth(se = FALSE)

grid.arrange(p1,p2,p3,p4, ncol=2,nrow=2)
```

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

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



When there is a snowfall(WT09=1), we get no attendance number, which means that snowfall definitely influences the attendance, or we can say snowfall infulences whether the baseball game would be held or not. As for other variables, like WT02(Heavy fog or heaving freezing fog, ice), WT03(Thunder), WT08(Smoke or haze), we actually cannot conclude directly from above, alghouth there are much more games when there is WT02 or WT 08 or WT09 equals 0.

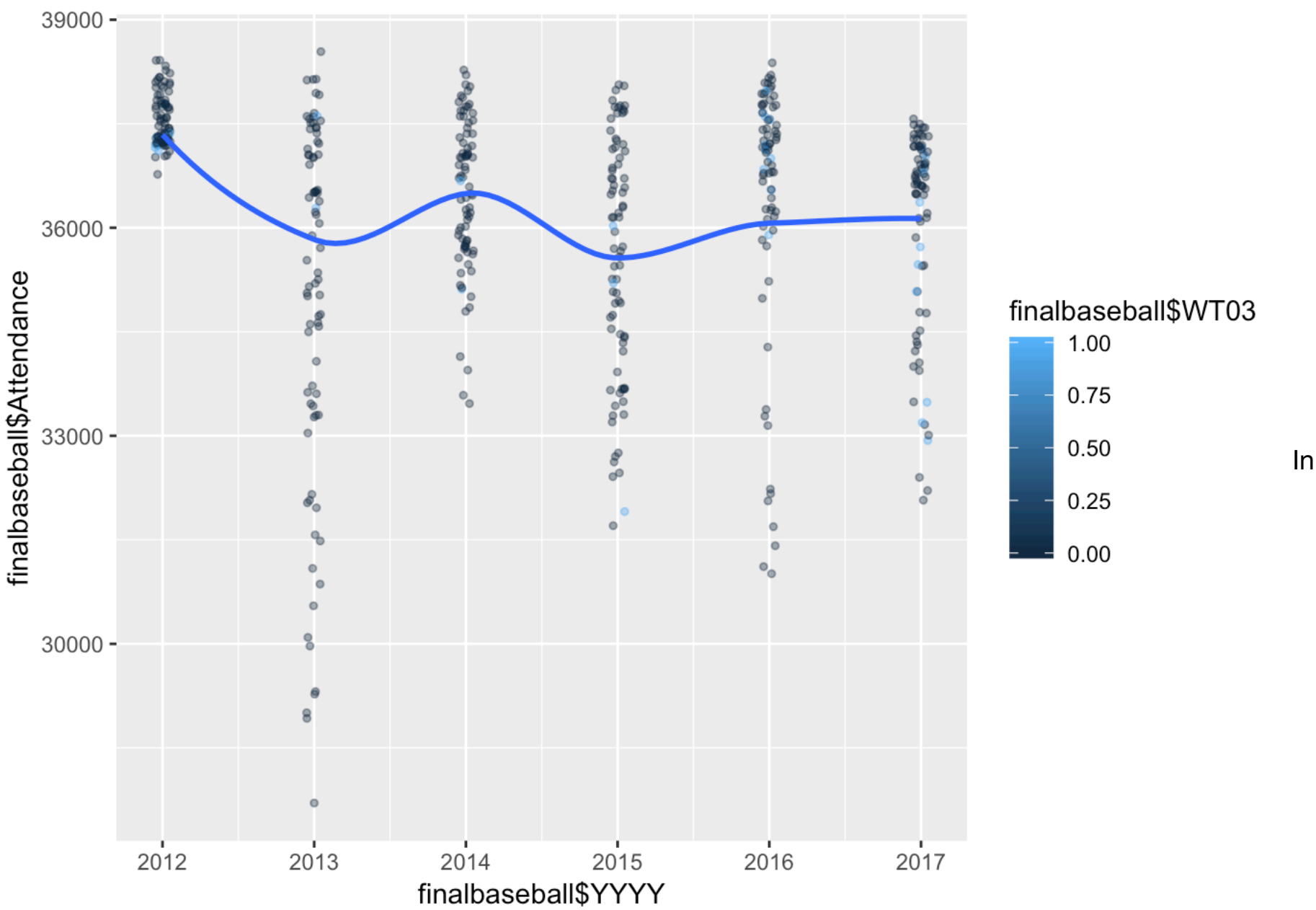
So maybe we can make comparison between different extreme weather. From the plots above, we can know that when WT02=1, the plots are more concentrated and the plots are more scattered while WT08=1. So we would say smoke or haze has more influence on attendance than other 3 weather types.

But we just cannot prove that there will be less attendance through visualization, on the contrast, the scatters distribution when WT=0 and WT=1 seems to be similar. Thus, more studies are necessary.

On the other hand, we find something interesting: Every year the distribution of the attendance numbers is different:

```
ggplot(data = finalbaseball,mapping = aes(y=finalbaseball$Attendance,x=finalbaseball$
YYYY,color=finalbaseball$WT03))+
          geom_jitter(width = 0.05,size=1, alpha=0.4)+
          geom_smooth(se = FALSE)
```

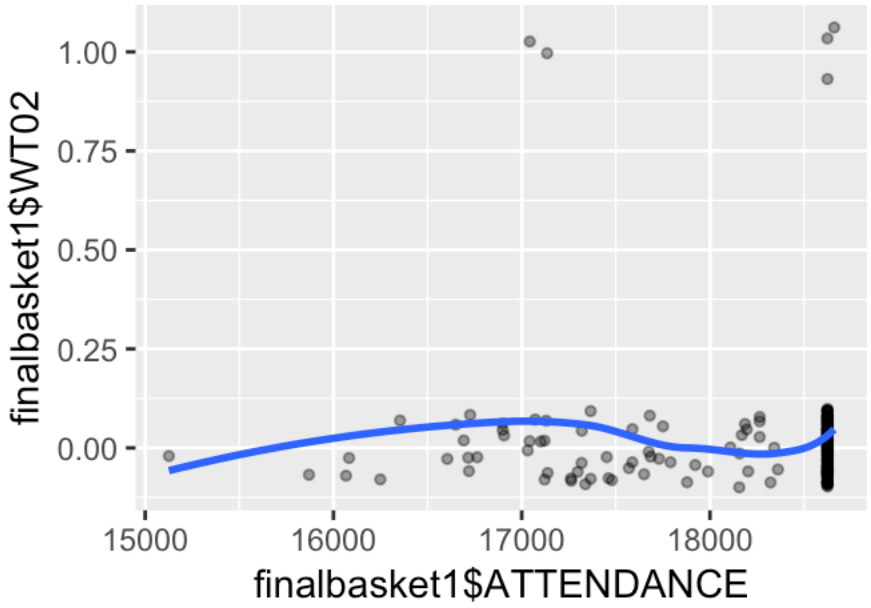
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

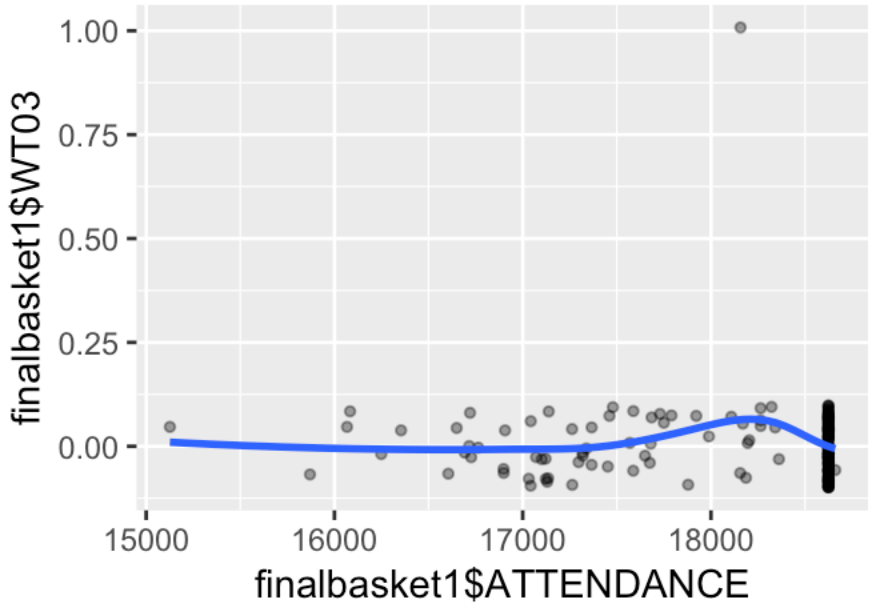


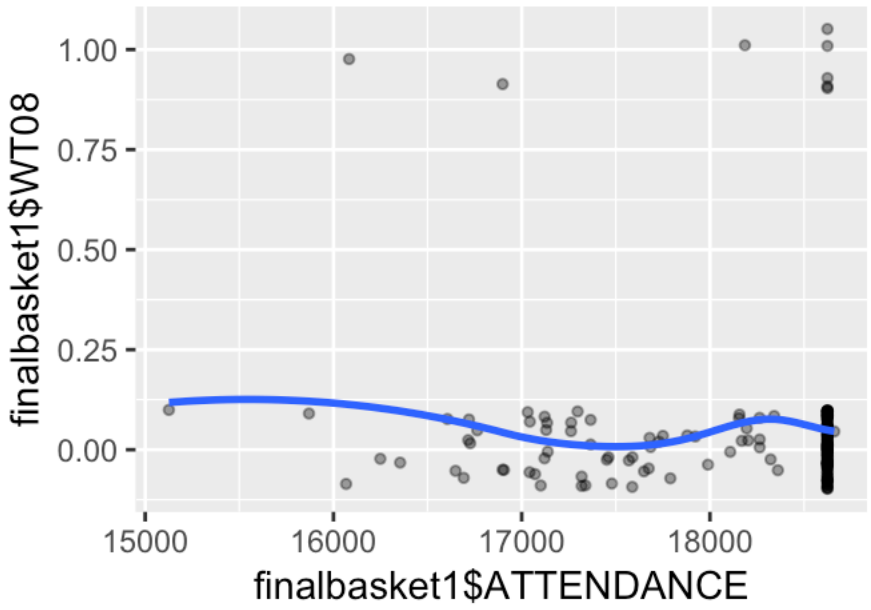
2012, most of the time the attendance number is more than 37000, which is unbelievable; In 2013, the attendance number is more scattered, which means the fans are less crazy. In 2014-2017, the situations are similar. This is so interesting since in 2012, Red Sox finished last in the five-team American League East, while in 2013, the Red Sox finished first in the American League East.

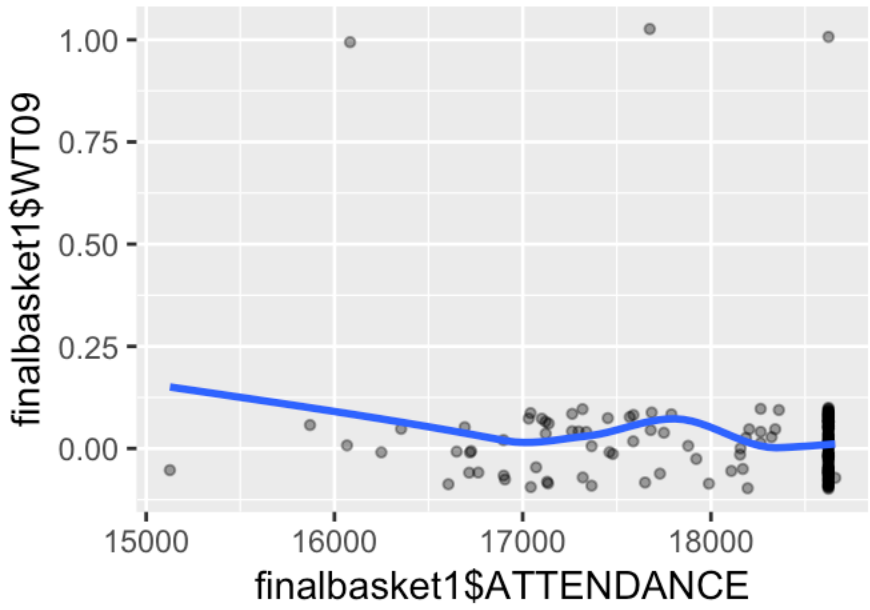
```
#In this part, the first thing we need to do is to remove the data in 2012 and 2017 s
ince this two years' attendance data are all 18624. It would be noisy if we keep them
in our analysis.
finalbasket1 <- filter(finalbasket,YEAR==2013|YEAR==2014|YEAR==2015|YEAR==2016)
require(gridExtra)
p1 <- ggplot(data = finalbasket1,mapping = aes(y=finalbasket1$WT02,x=finalbasket1$ATT
ENDANCE))+
  geom_jitter(height=0.1,size=1, alpha=0.4)+
  geom_smooth(se = FALSE)
p2 <- ggplot(data = finalbasket1,mapping = aes(y=finalbasket1$WT03,x=finalbasket1$ATT
ENDANCE))+
  geom_jitter(height=0.1,size=1, alpha=0.4)+
  geom_smooth(se = FALSE)
p3 <- ggplot(data = finalbasket1,mapping = aes(y=finalbasket1$WT08,x=finalbasket1$ATT
ENDANCE))+
  geom_jitter(height=0.1,size=1, alpha=0.4)+
  geom_smooth(se = FALSE)
p4 <- ggplot(data = finalbasket1,mapping = aes(y=finalbasket1$WT09,x=finalbasket1$ATT
ENDANCE))+
  geom_jitter(height=0.1,size=1, alpha=0.4)+
  geom_smooth(se = FALSE)

grid.arrange(p1,p2,p3,p4, ncol=2,nrow=2)
```

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



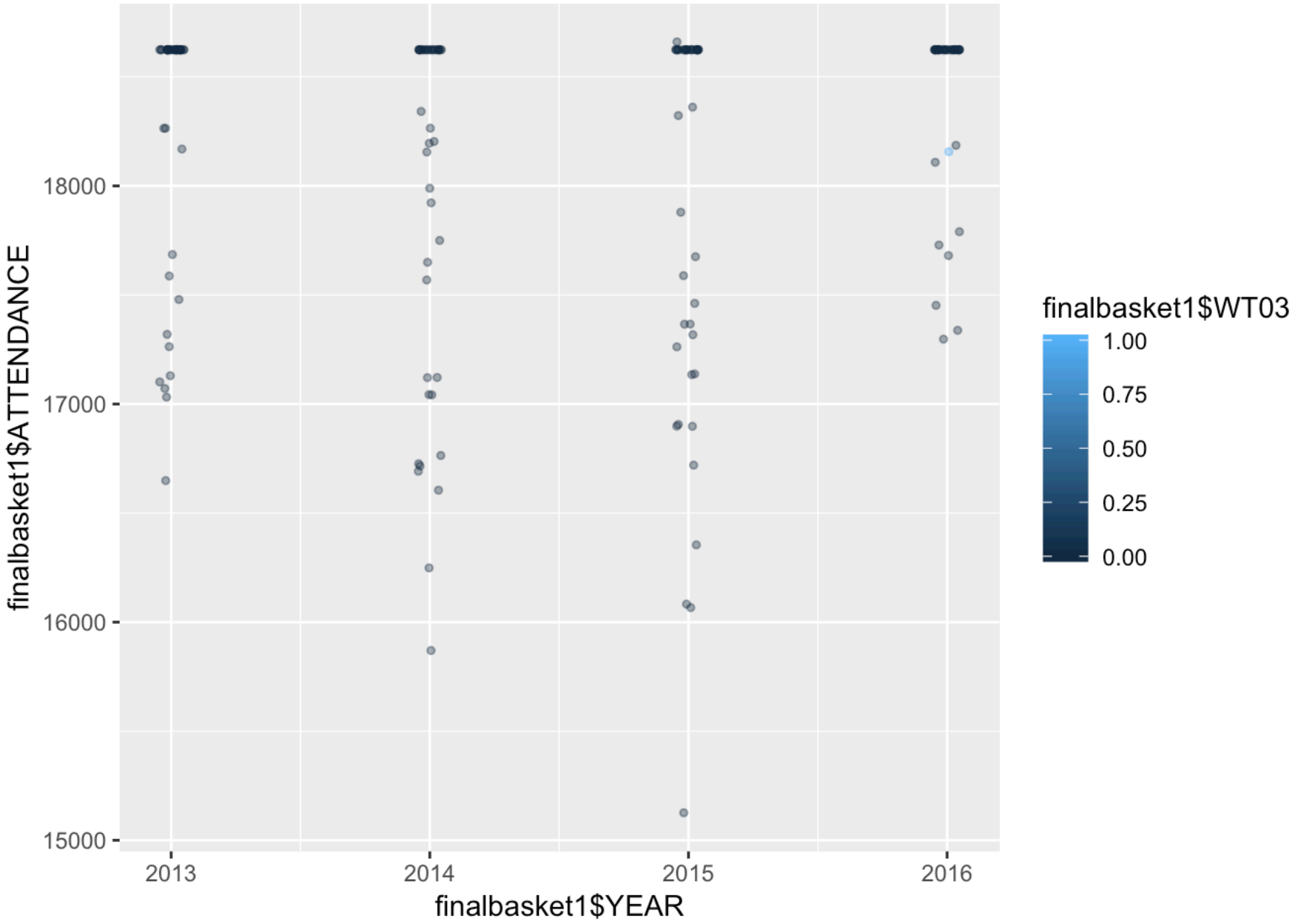






We'd better say we don't know the relation between the extreme weather types and the attendance since the sample size is too small.

```
#Then we do the same thing for basketball data:
ggplot(data = finalbasket1,mapping = aes(y=finalbasket1$ATTENDANCE,x=finalbasket1$YEA
R,color=finalbasket1$WT03))+
          geom_jitter(width = 0.05,size=1, alpha=0.4)
```



It seems that basketball game attendance every year is similar. This is different from baseball.

Conclusion:

Average wind speed vs attendance： Accoridng There's no significant relationship between basketball and average wind speed. We found this is mainly because basketball games are held indoor. Audience are less influenced by wind speed.

Average temperature vs attendance: We can see the attendance trend of the interval between 40 and 70 are going up showing that there might be some positive relationship between temperature and attendance. Later on, the weather does not affect the attendance a lot. For basketball, we are unable to see the clear trend between the temperature and attendance using all years long data.

Averave precipitaion vs attendance rate: By comparing the two ggplots, we can see the effect of precipitation on baseball game attendance is much more than that on basketball game attendance. The reason for this maybe is that baseball games are held outside but basketball games are hled in gyms. Besides, the tendency of baseball game attendance decreases as the degree of precipitation increases although there exist unavoidable fluctuations in it.

Weather types vs attendance rate: From the plots above, it's even more obscure to find the relation between weather type and attendance since we cannot find the difference between distribution of WT=0 and distribution of WT=1 with such a little spots of WT=1. Also even the weather type is extreme, there is also

some spots showing that many people attended the games.