615Midtermmmmmmmm

Kaiyu Yan,Si Chen, Wenjia Xie, Siwei Hu

October 17, 2018

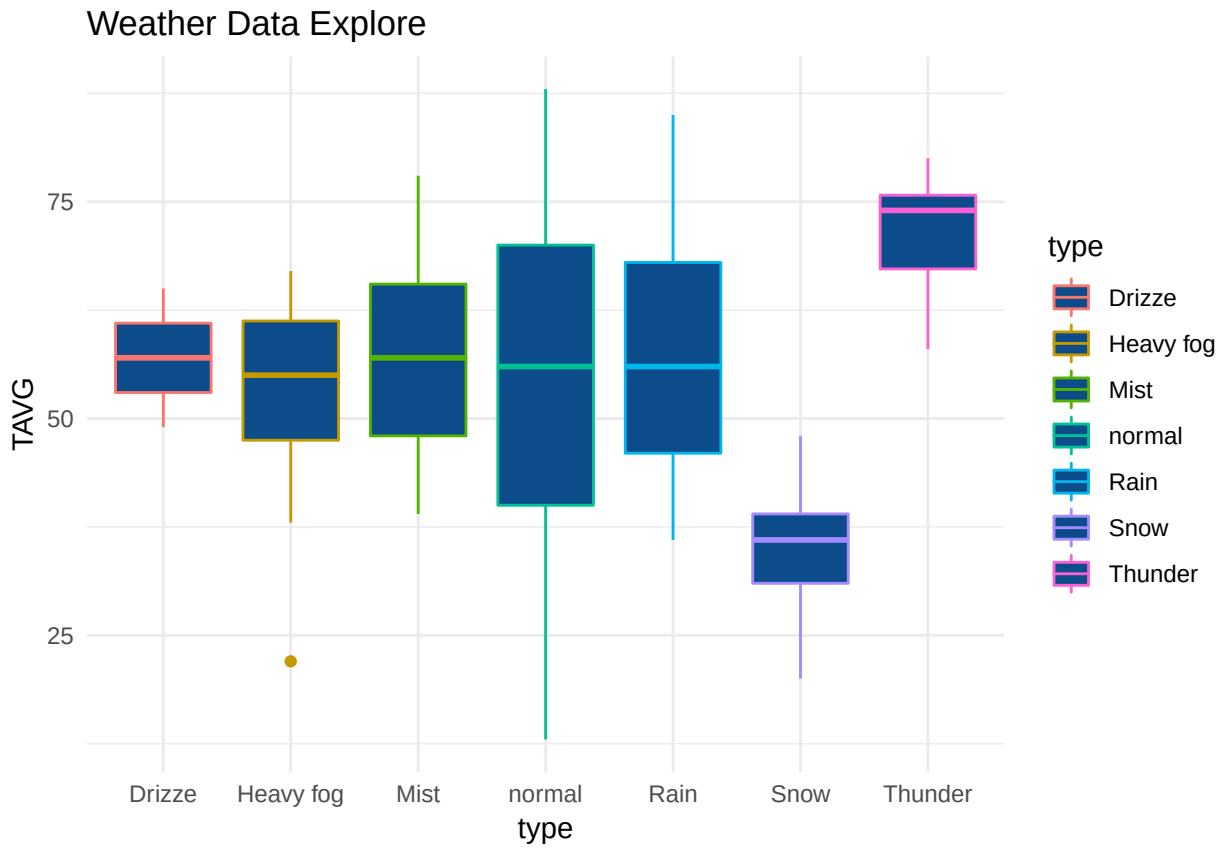
We send email to noaa.com and asked for their weather data. They send us a Data with 18 variables which are different types of weather and max/min temperature. We choose 7 different types of weatherand changes their jorgon to normal words like"snowy" "fog" to help us understand what happen.We make a new column called "type" and then we build a double for loop in order to get weather type everyday.

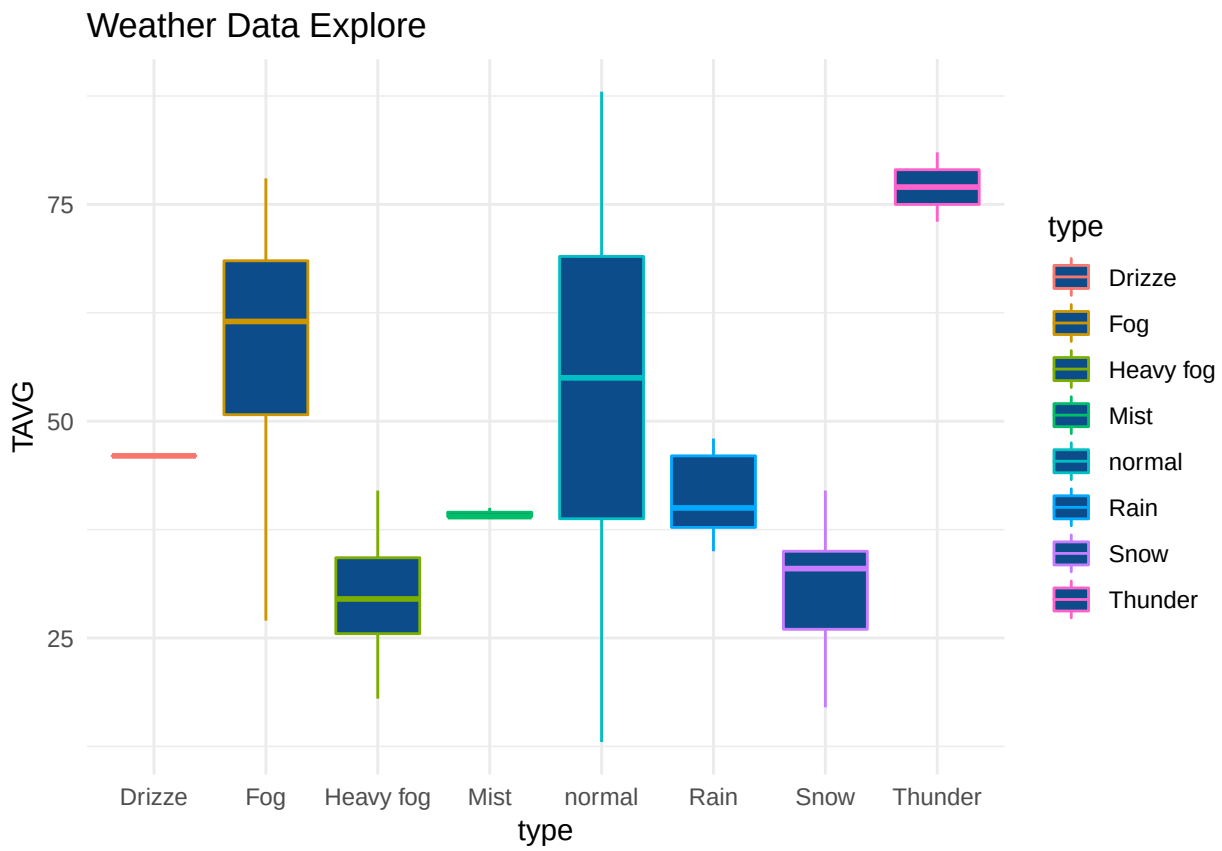
```
##### new
##### Weather Data #####

#Weather data cleaning
weather_data <- read.csv("weather data.csv")
colnames(weather_data)[colnames(weather_data)=="DATE"] <- "Date"
weather_data$Date <- as.Date(weather_data$Date,"%Y-%m-%d")
#Select certain weather type to do analysis
typelist = c("Fog","Mist","Drizze","Rain","Snow","Thunder","Heavy fog")
type_code = c("WT01","WT13","WT14","WT16","WT18","WT03","WT02")
weather_data$type<-NA

for (i in 1:length(typelist)) {
  colnames(weather_data)[which(colnames(weather_data)==type_code[i])] = typelist[i]
}

weather_data[is.null(weather_data)] <- NA

##Create a new column that contain certian weather type for each day
for (m in 1:dim(weather_data)[1]) {
  t<-0
  for (n in 1:length(typelist)) {
    if (is.null(weather_data[m,typelist[n]])) {
      weather_data[m,typelist[n]] = NA
    }
    if (!is.na(weather_data[m,typelist[n]])) {
      weather_data[m,"type"] =  typelist[n]
      t<-t+1
    }
  }
  if(t==0)
    weather_data[m,"type"] = "normal"
}
```

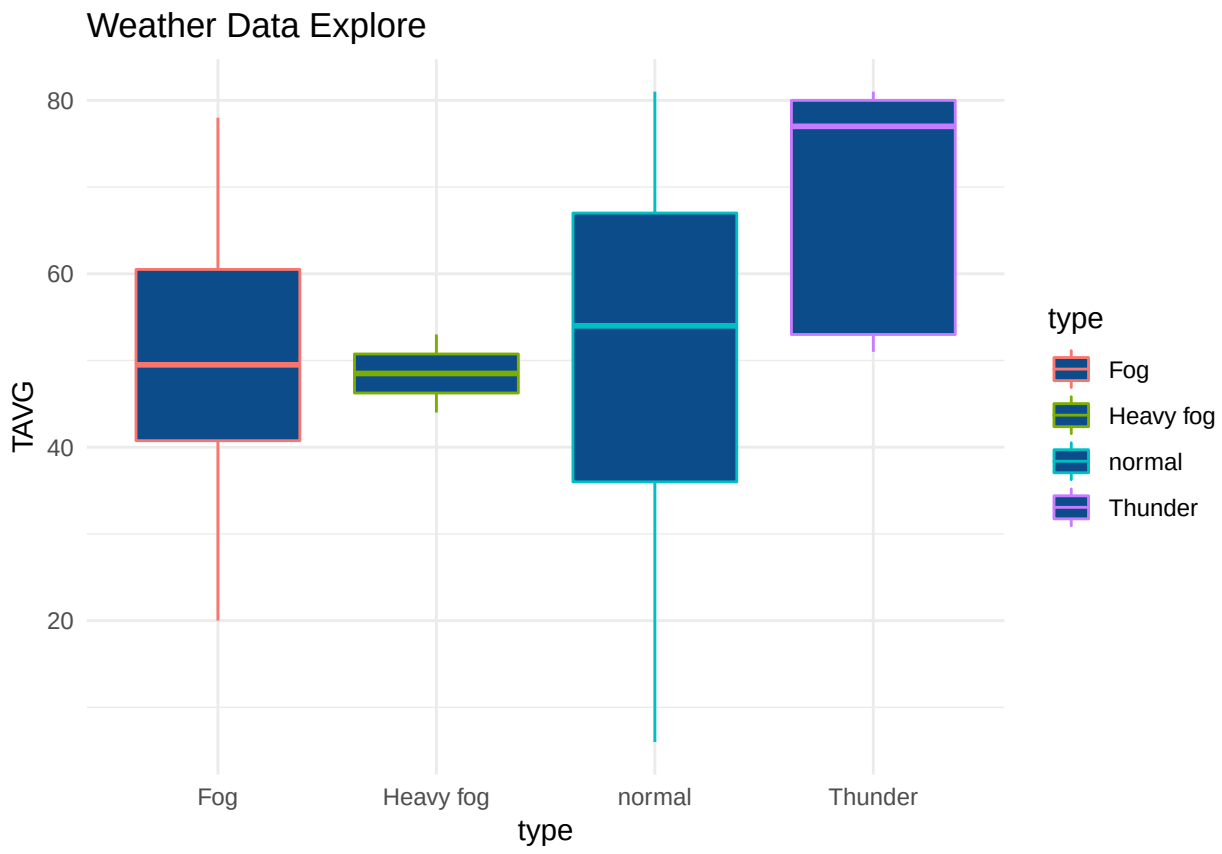
Here is the relationship between Weather and Average Temperature. Here we want to use these plots to understand weather type and help to prove the relationship between Weather type and attendence.

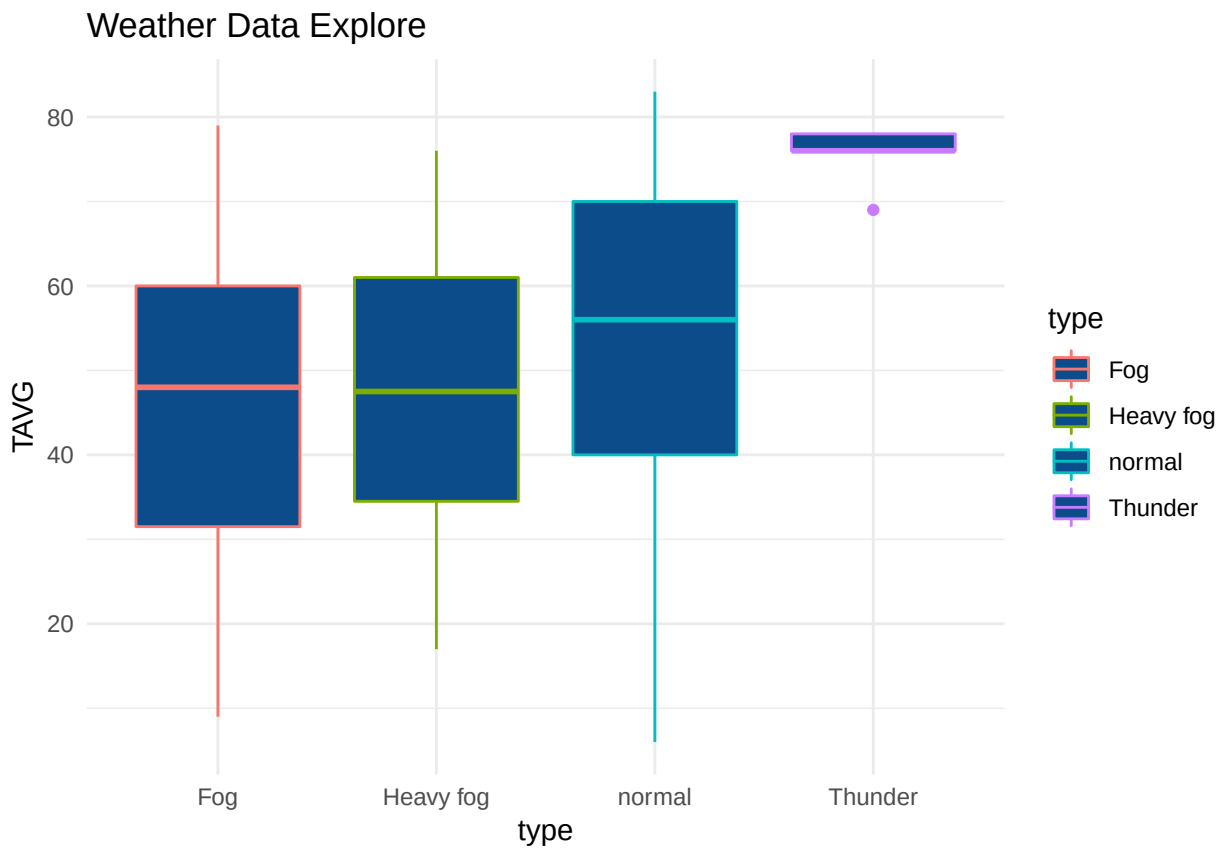
```
year<-c(2012:2017)
#Plot weather data
for (i in 1:6){
weather_year<-filter(weather_data,str_detect(Date,as.character(year[i])))
pic<-ggplot(data = weather_year) +
  aes(x = type, y = TAVG, color = type) +
```

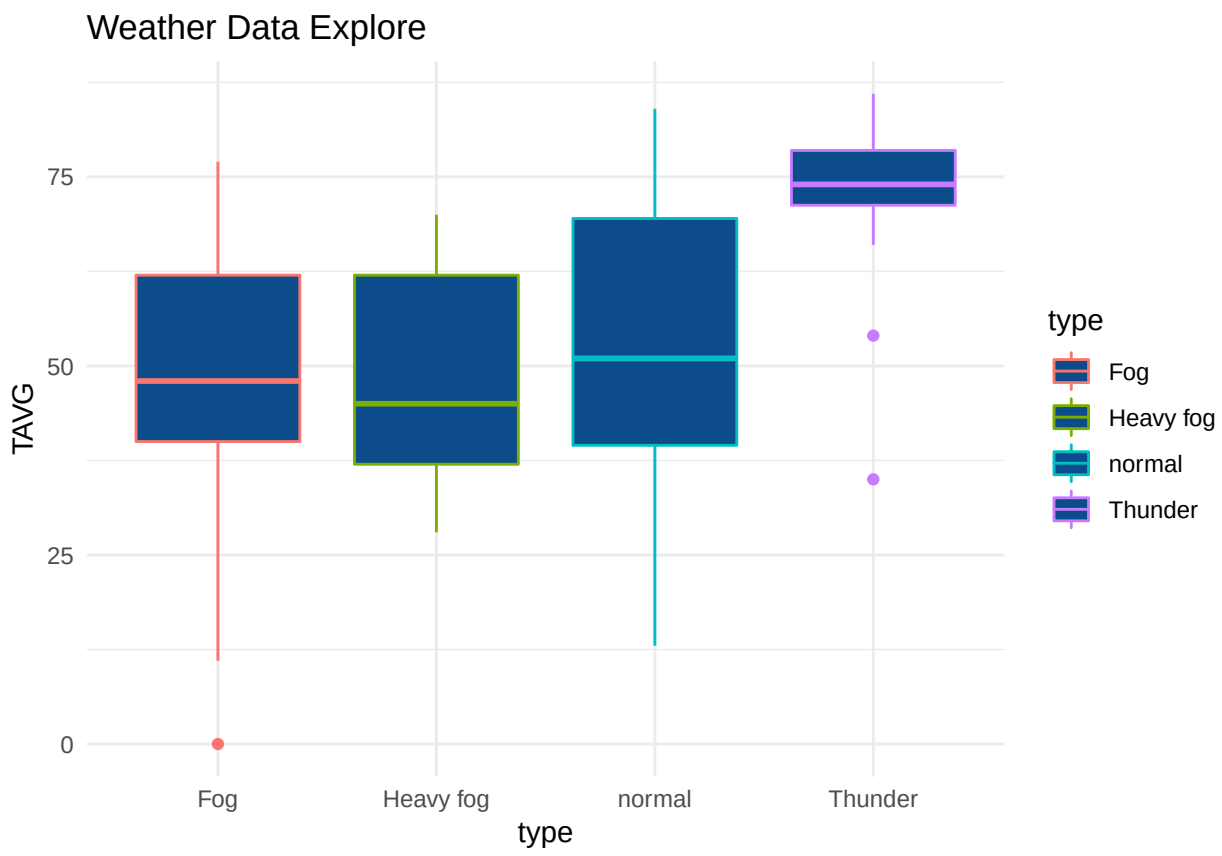
```
  geom_boxplot(fill = "#0c4c8a",notch = FALSE) +
  theme_minimal()+
  ggtitle("Weather Data Explore ")
print(pic)
}
```

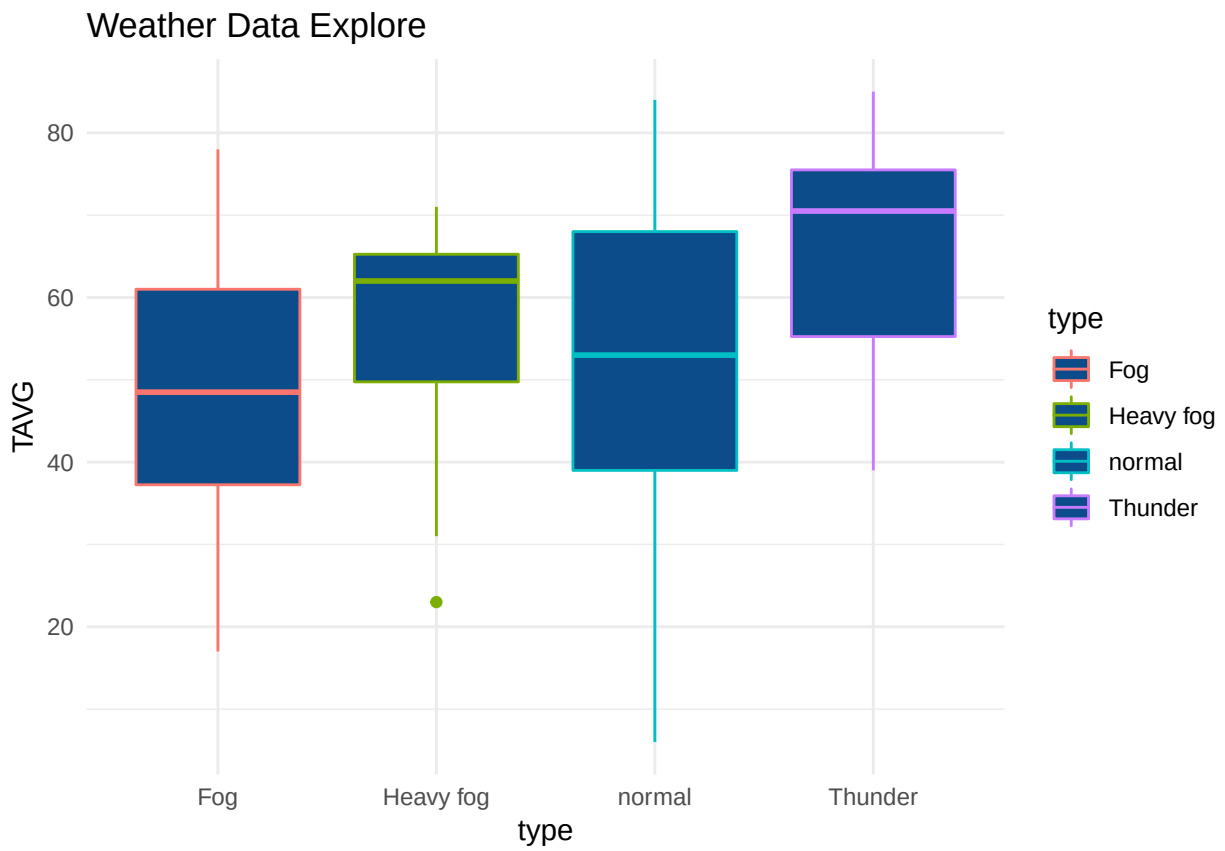



Weather Data Explore
TAVG
75
50
25
Drizze
Fog
Heavy fog
Mist
normal
Rain
Snow
Thunder
type
type
Drizze
Fog
Heavy fog
Mist
normal
Rain
Snow
Thunder


Weather Data Explore
80
60
40
20
TAVG
Fog
Heavy fog
normal
Thunder
type
type
Fog
Heavy fog
normal
Thunder


Weather Data Explore
80
60
40
20
TAVG
Fog
Heavy fog
normal
Thunder
type
type
Fog
Heavy fog
normal
Thunder


Weather Data Explore
TAVG
75
50
25
0
Fog
Heavy fog
normal
Thunder
type
type
Fog
Heavy fog
normal
Thunder



We got Basketball data from "baskerball-reference". We write code to grab data from their website from 2012 to 2017 season. In this chuck i put code that how we grab data from website and how we build the dataframe. Since we finish the data frame, we export it to be a .csv file for convenience.

```
# library(xml2)
# library(rvest)
# library(tidyverse)
# site1 <- "https://www.basketball-reference.com/leagues/NBA_"
# site2 <- "_games-"
# site3 <- ".html"
# year <- c(2013:2018)
# month <- c("january","february","march","april","may","june","october","november","december")
# month_index<-c(1:9)
# year_index<-c(1:6)
# name1<-c()
# name2<-c()
# total.date<-c()
# total.attdence<-c()
# for (i in year_index){
#   for (j in month_index){
#     site<-paste(site1,year[i],site2,month[j],site3,sep="")
#     webpage<-read_html(site)
#     name1<- webpage %>% html_nodes('.left:nth-child(1)') %>% html_attrs()
#     name2<- webpage %>% html_nodes('.center+.right') %>% html_text()
#     total.date <- c(total.date,name1)
#     total.attdence <- c(total.attdence,name2)
#     j<-j+1
```

```
#   }
#   i<-i+1
# }
# total.date1 <- t(data.frame((total.date)))[,4]
# total.attdence1 <- data.frame(total.attdence)
# b_data<-cbind(total.date1,total.attdence1)[-1,]
# colnames(b_data)<-c("date","attendance")
# rownames(b_data)<-rep(1:7934)
# host<-substr(b_data$date, 10, 12)
# b_data$date<-substr(b_data$date,1,8)
# b_data3<-cbind(b_data,host)
# write.csv(b_data3,'basketball.csv')
```

Then we grab host variable with string "BOS" to get date when Celitics play in their host "TD Garden". Because date we grab have one more 0 like"201201010". So we use substring to cut out only 8 digitals and convert date type to "yyyy-mm-dd" in order to join with weather data.

```
##### Basketball Data Only #####

#Data Cleaning for basketball data
Basketball<-read.csv("basketball.csv")
Homegame <- filter(Basketball,str_detect(host,"BOS"))
Homegame$date <- substring(Homegame$date, 1,8)
#Convert original date type
Homegame$date <- as.Date(Homegame$date,"%Y%m%d")
names(Homegame)[2]<-paste("Date")
names(Homegame)[3]<-paste("Attendance")
```

We got Baseball data from "baseball-reference". We write code to grab data from their website from 2012 to 2017 season. In this chuck i put code that how we grab data from website and how we build the dataframe. Since we finish the data frame, we export it to be a .csv file for convenience. Btw, we grab baseball data each year because the baseball season is from April to October.

```
# site1 <-"https://www.baseball-reference.com/teams/BOS/"
#site2 <- "-schedule-scores.shtml"
# year <- c(2013:2018)
# month_index<-c(1:9)
# year_index<-c(1:6)
# name1<-c()
# name2<-c()
# total.date<-c()
# total.attdence<-c()
# for (i in year_index){
#   for (j in month_index){
#     site<-paste(site1,year[i],site2,sep="")
#     webpage<-read_html(site)
#     name1<- webpage %>% html_nodes('.left:nth-child(1)') %>% html_attrs()
#     name2<- webpage %>% html_nodes('.center+.right') %>% html_text()
#     total.date <- c(total.date,name1)
#     total.attdence <- c(total.attdence,name2)
#     j<-j+1
#   }
#   i<-i+1
# }
#
```

```
# total.date1 <- t(data.frame((total.date)))[,4]
# total.attdence1 <- data.frame(total.attdence)
# b_data<-cbind(total.date1,total.attdence1)[-1,]
# colnames(b_data)<-c("date","attendance")
# rownames(b_data)<-rep(1:7934)
# host<-substr(b_data$date, 10, 12)
# b_data$date<-substr(b_data$date,1,9)
# b_data3<-cbind(b_data,host)
#
# class(b_data3$attendance)
# write.csv(b_data3,'Baseball.csv')
```

WE select several variable include Date, Attendance, year and etc. Since the date of base is like "Sunday,04,11", we seperate "Sunday","04","11" to three different columns and merge year,month and day to be a new date variable, use filter function to get host information. We spend a lot of time to change this date type to become same as the date type in weather data. Final we get date type like "yyyy-mm-dd".

```
####### Baseball Data Only #######
#Baseball data cleaning
Baseball <- read.csv("Baseball.csv")
Clean_1 <- select(Baseball, Date, Tm:Opp, Attendance,Year)
#Convert original date type
Clean_2 <- data.frame(do.call('rbind', strsplit(as.character(Clean_1$Date),',',fixed=TRUE)))
Clean_2 <- data.frame(do.call('rbind', strsplit(as.character(Clean_2$X2),' ',fixed=TRUE)))
Clean_1 <- merge(Clean_1,Clean_2,by = 0)
Clean_3<- filter(Clean_1, !str_detect(Var.5,"@"))
Clean_4 <- filter(Clean_3,str_detect(Tm,"BOS"))
Clean_5 <- select(Clean_4, 6,7,9,10)
Clean_5$X2 <- match(Clean_5$X2,month.abb)
Baseball_All<- unite(Clean_5,Date,2:4,sep = "-",remove = TRUE)
Baseball_All$Date <- as.Date(Baseball_All$Date,"%Y-%m-%d")
```

We use inner_join to get weather data only when these two teams played in their host. After combine weather ,baseball and baketball, we divided them onto different tables by years in order to build ggplot easier.

```
##### Basketball Join Weather ######
Celtics_All <- inner_join(Homegame,weather_data,by = "Date", match = all)
Celtics_All <- select(Celtics_All,Date,Attendance,TAVG,type)
Celtics_All$Attendance <-  as.numeric(as.character(Celtics_All$Attendance))

#Full baseball data for each year
for(i in 2012:2017) {
  assign(paste("Celtics",i,sep="_"),filter(Celtics_All,str_detect(Date,paste(i))))

}
```

```
##### Baseball Join Weather #####
Redsox_All <- inner_join(Baseball_All,weather_data,by = "Date", match = all)
Redsox_All <- select(Redsox_All,Date,Attendance,TAVG,type)
Redsox_All$Attendance <-  as.numeric(as.character(Redsox_All$Attendance))

#Full baseball data for each year
for(i in 2012:2017) {
  assign(paste("Redsox",i,sep="_"),filter(Redsox_All,str_detect(Date,paste(i))))
```

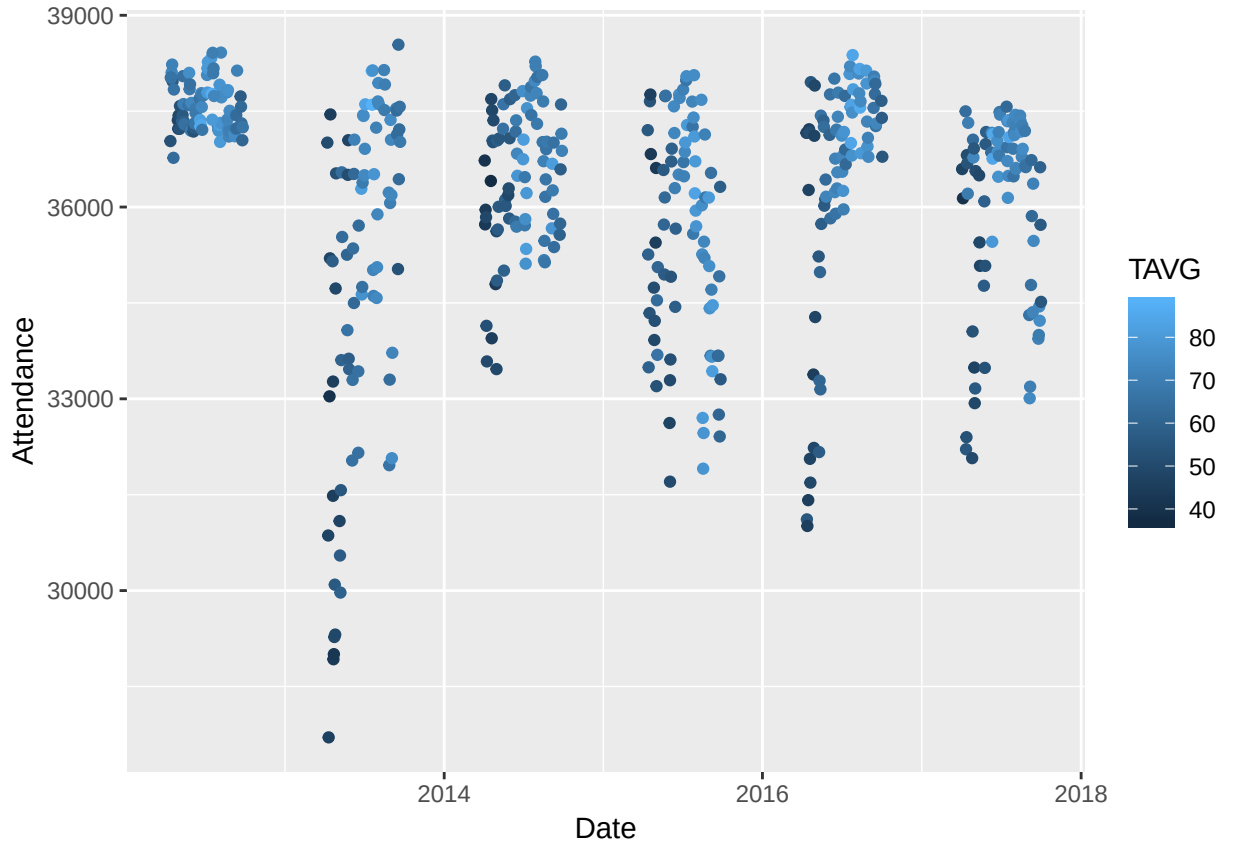
```
}
```

After gathering the attendence of baseball and weather data together, we made several plots to view the relations between attendence and average temperature.

From the first two plots, we gain an overall picture of the data. Generally, when temperature gets higher, more people are willing to watch the game. This can be seen from the first graph that darker points(which indicate lower temperature) are normally distrubuted at a lower place in the plot(which indicate lower attandence). The second plot also lead to the similar conclusion, while at the same time we can also see that temperature is a more important factor when compared with weather type.

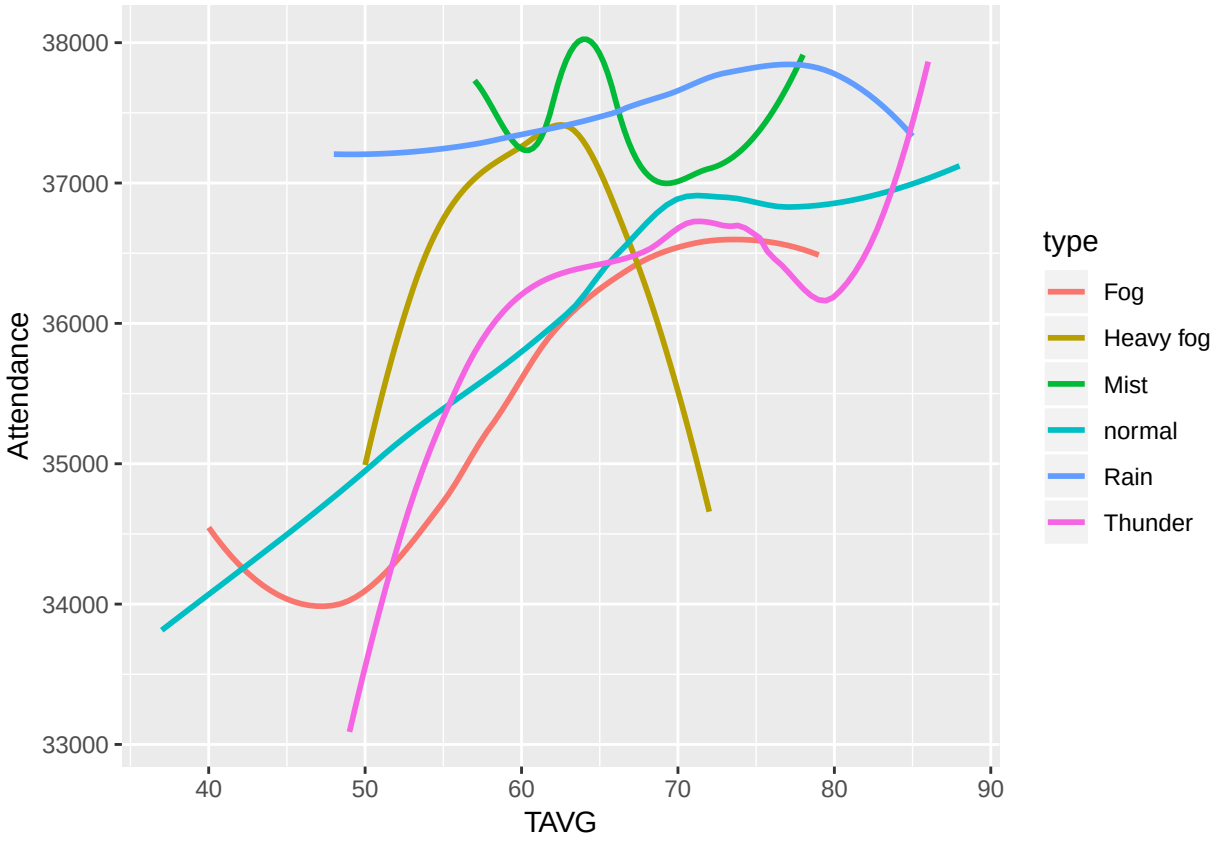
In the following six plots, we plot the scatter plots to see the relations between temperature and attandence for every year.Different colours are also used to show the influence of different weather types.An interesting discovery is that on thunder, people are even more willing to watch the game in some years. For this phenomenon, we think it might because thunder is often occured in summer,when the temperature is relatively high. Thus the high temperature may be the main cause rather than thunder.

```
ggplot(data = Redsox_All,aes(x = Date, y = Attendance, color = TAVG))+geom_point( )
```



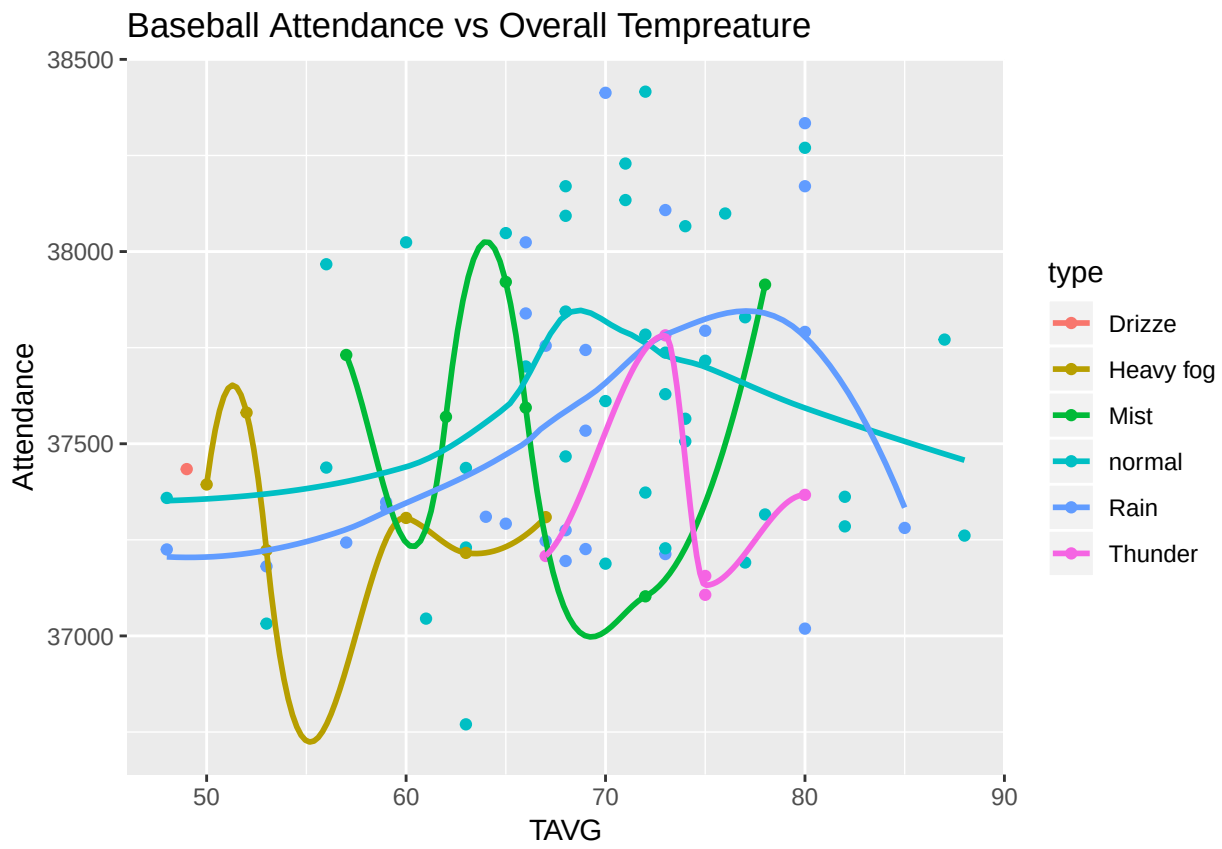

```
ggplot(data = Redsox_All) +
  geom_smooth(mapping = aes( x = TAVG, y = Attendance, color = type), se = F)
```

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

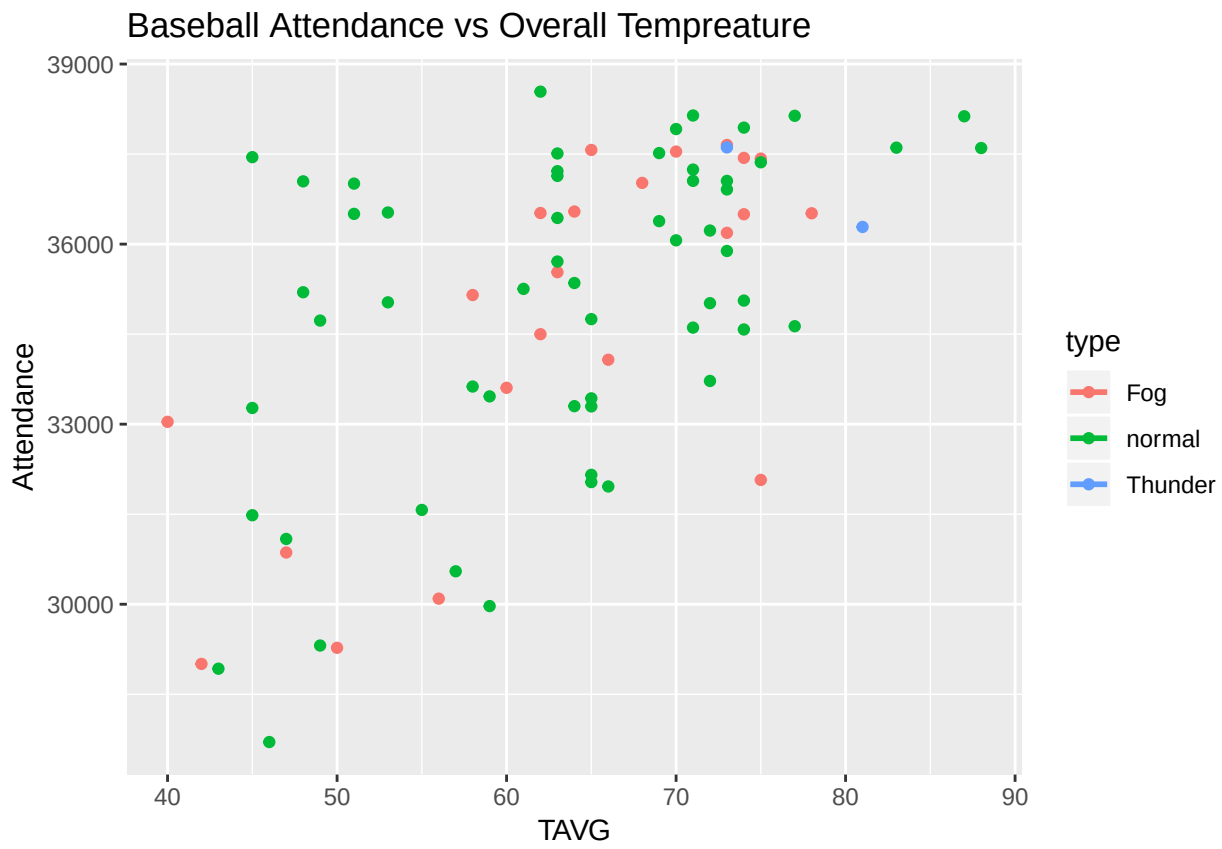



```
Dataset= list(Redsox_2012,Redsox_2013,Redsox_2014,Redsox_2015,Redsox_2016,Redsox_2017)

for(i in 1:6){
  p<- ggplot(data = Dataset[[i]],aes(x = TAVG, y = Attendance,  color = type))+
    geom_point()+geom_smooth(se=F)+
    ggtitle("Baseball Attendance vs Overall Tempreature ")
    print(p)
    }
```

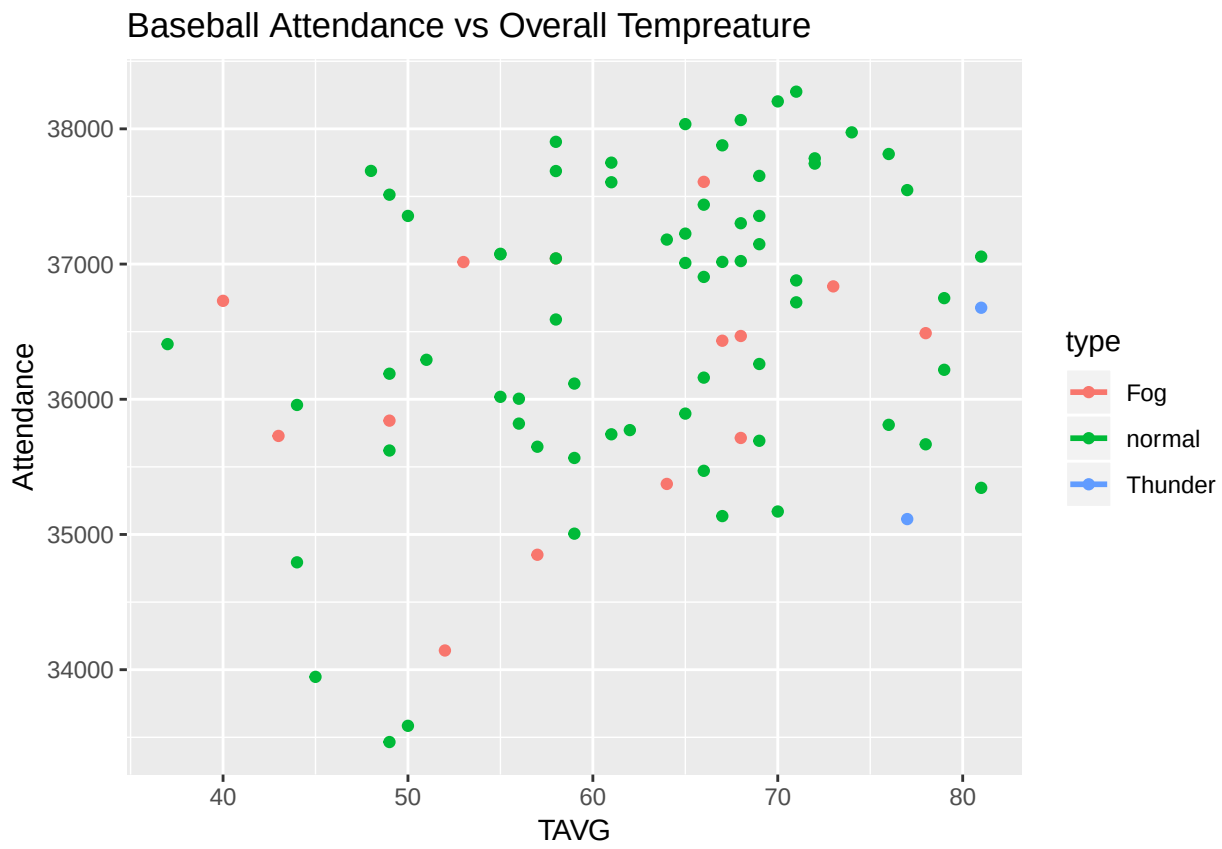
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



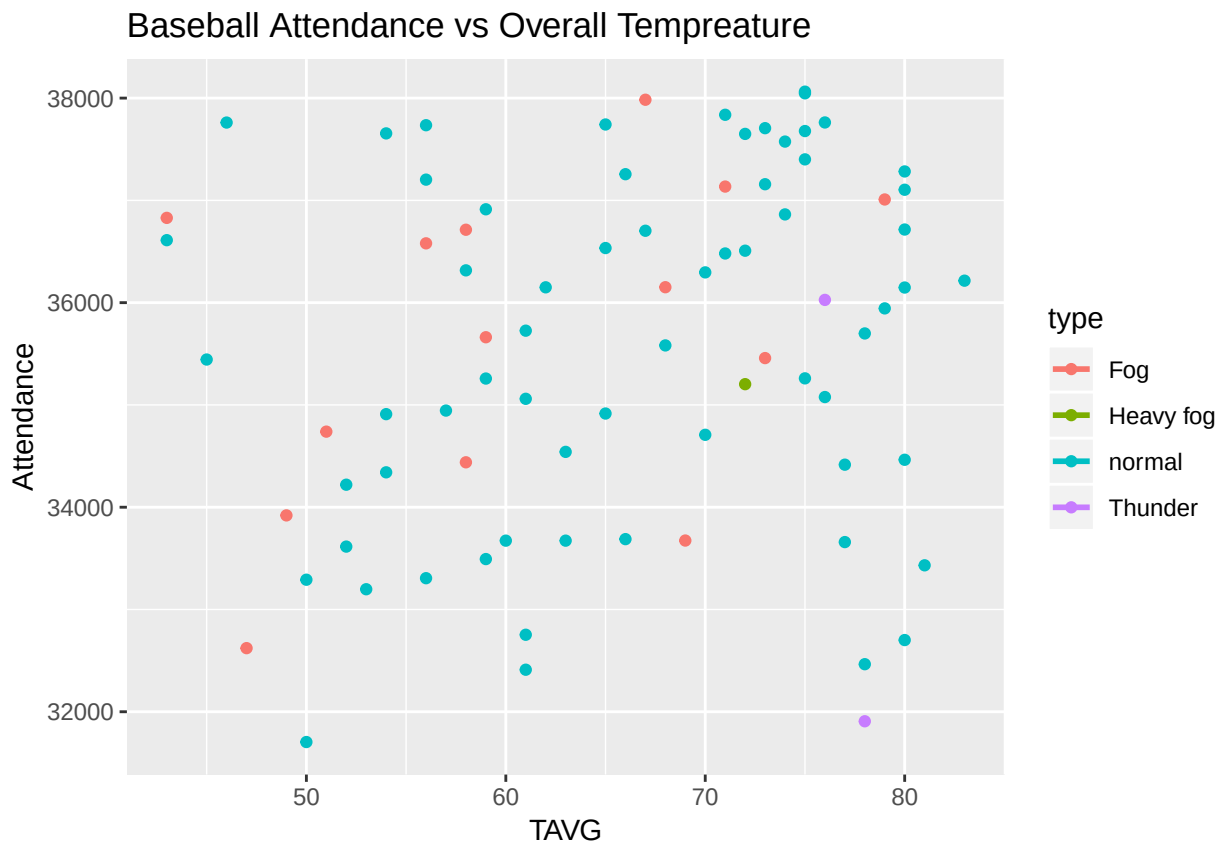

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



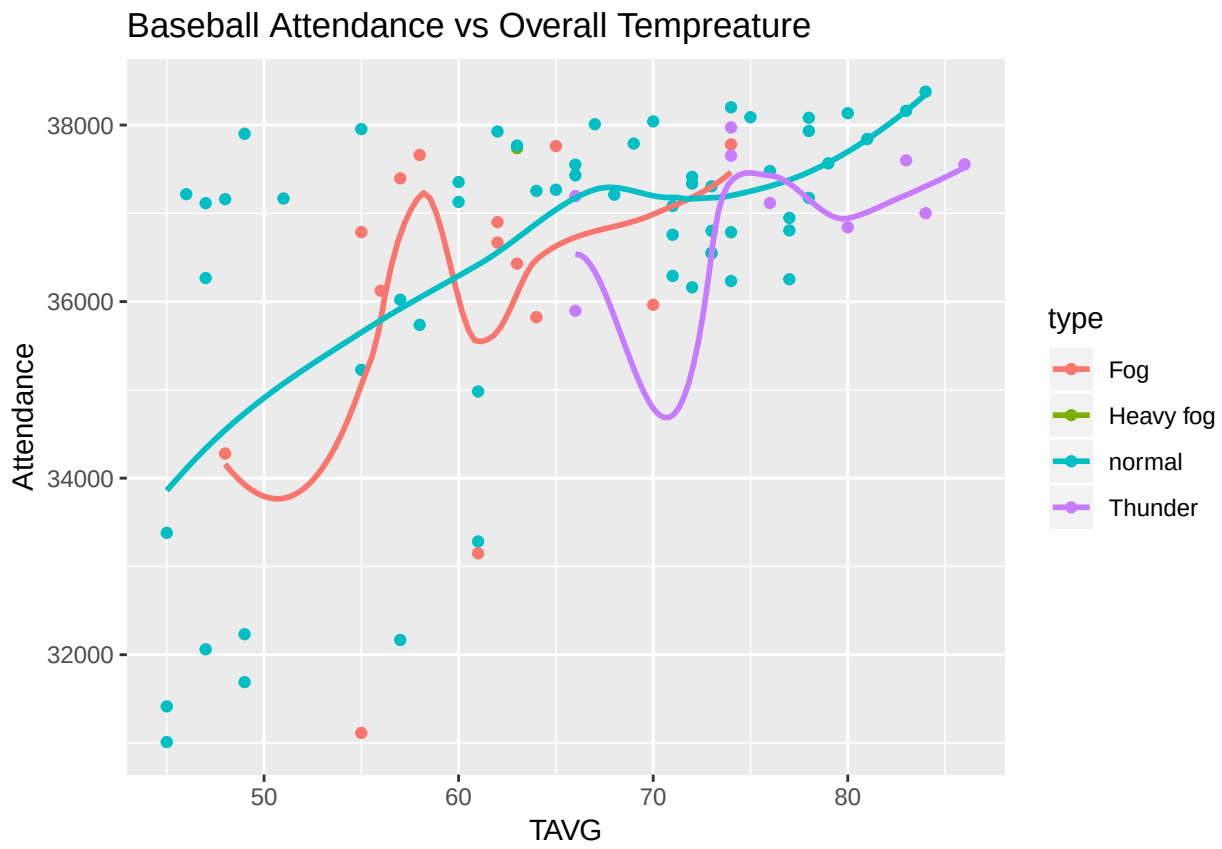

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



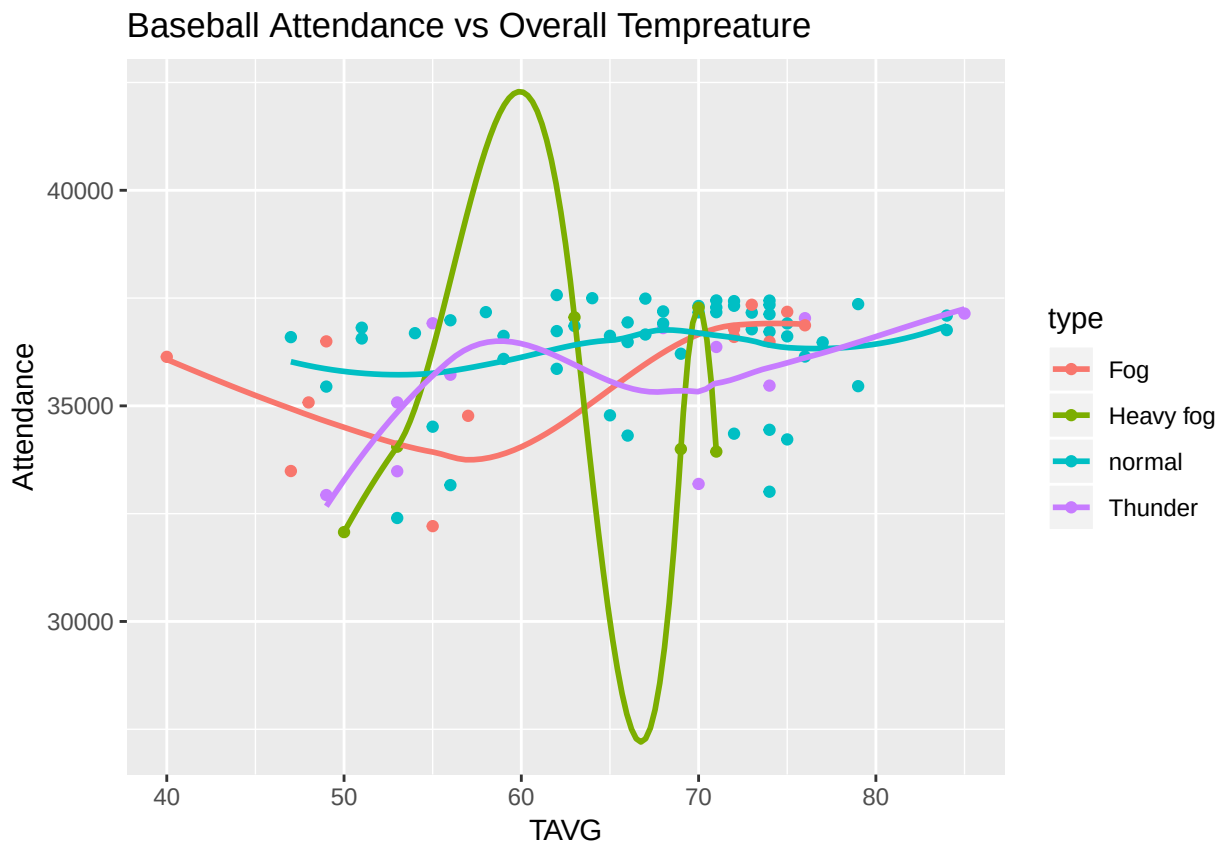

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




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



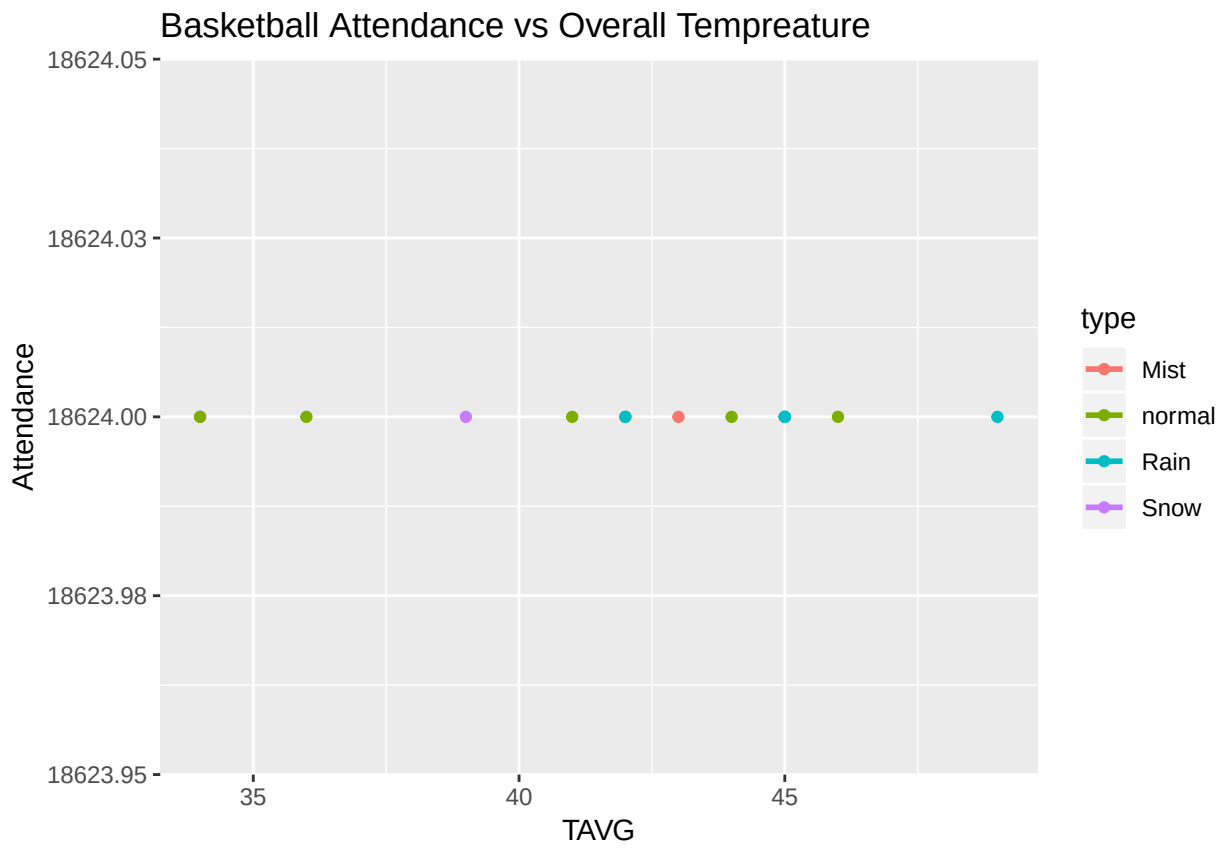

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



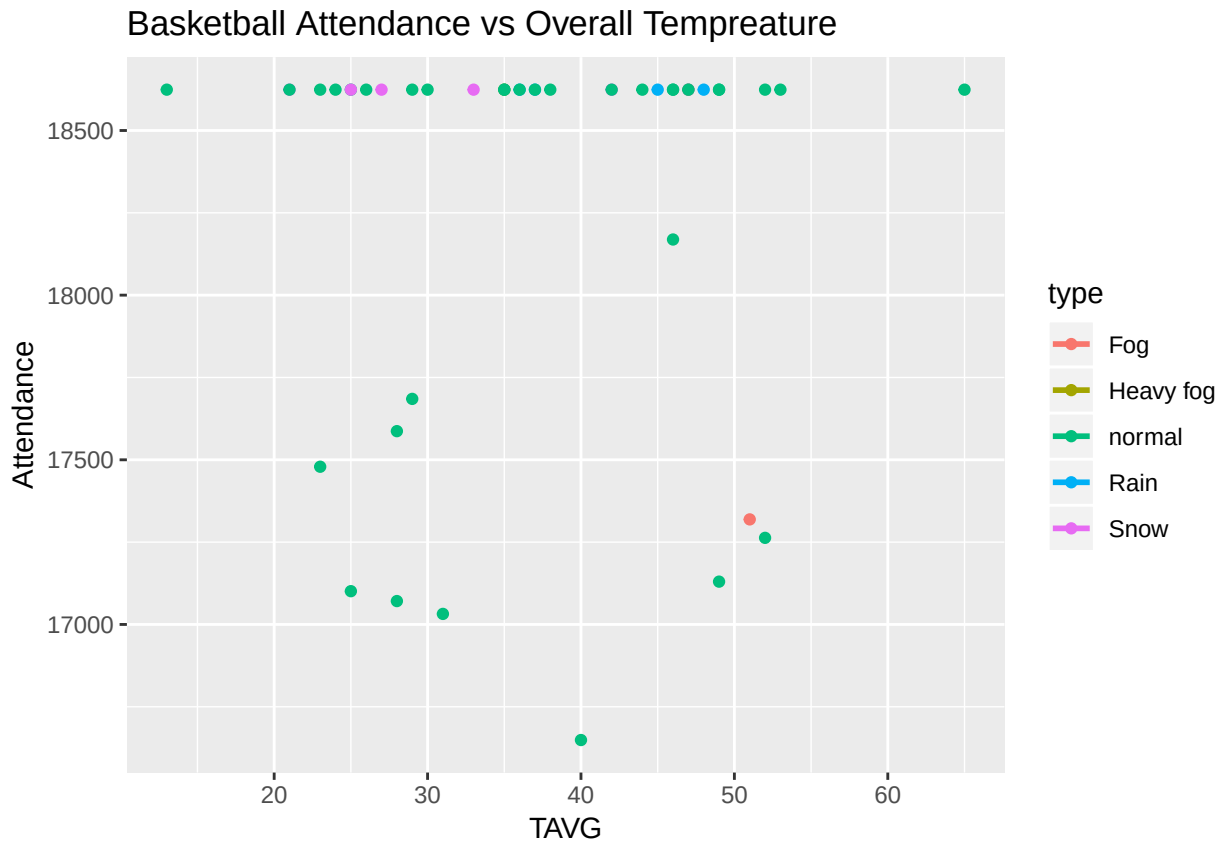
We built point graph and smooth line on the ggplot, set x-axis = Average temperature and y-axis = attendance and use different colors to show the type of that day's weather. We add one more layer to draw smooth line to show how different weather affect the attendence. From the data, basketball was not affected too much by weather especially the data in 2012 which all are 18624. The reason is that it has stadium. Althogh we try to draw smooth line for all year, but actually some year 's data structure are not fit to draw a smooth line. So some ggplot only show points on it.

```
Dataset2= list(Celtics_2012,Celtics_2013,Celtics_2014,Celtics_2015,Celtics_2016,Celtics_2017)


for(i in 1:6){
  q <- ggplot(data = Dataset2[[i]],aes(x = TAVG, y = Attendance,  color = type))+
    geom_point()+geom_smooth(se=F)+
     ggtitle("Basketball Attendance vs Overall Tempreature ")

    print(q)


    }
```

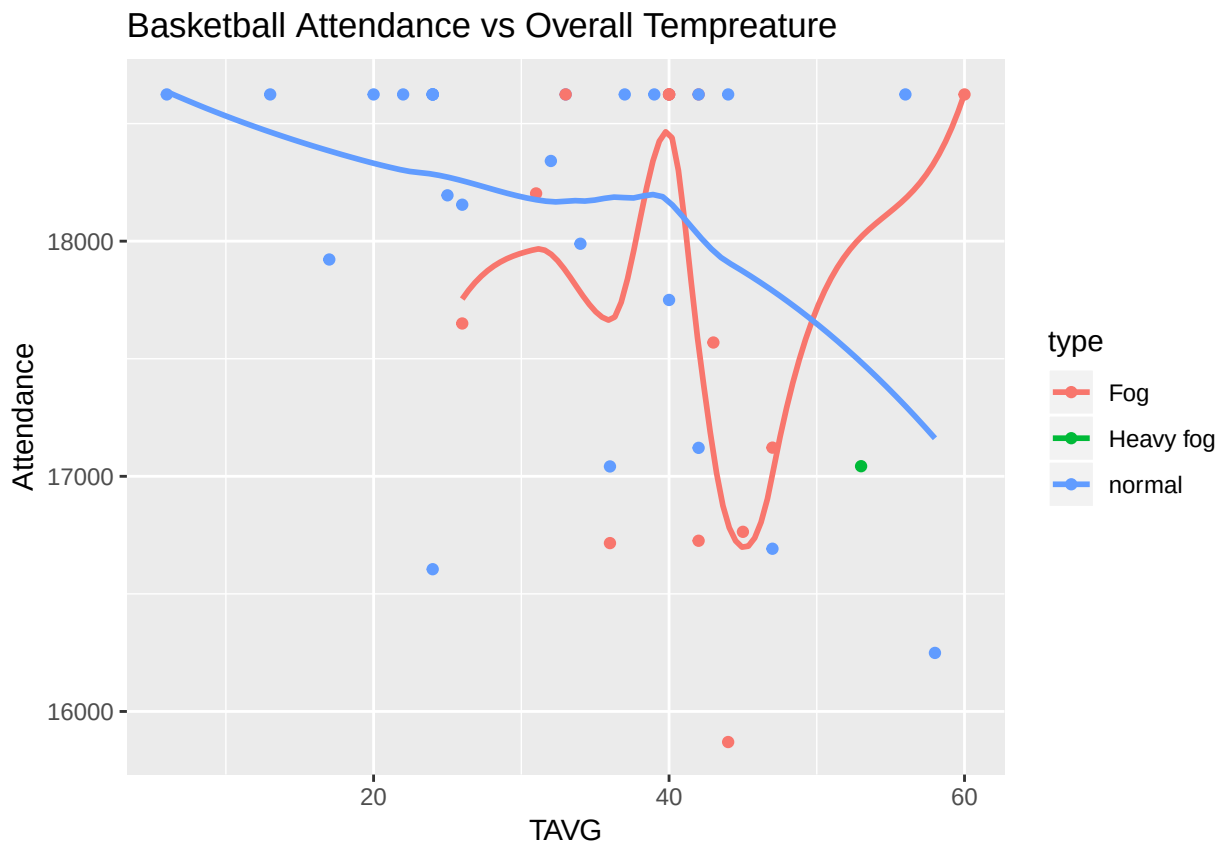
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



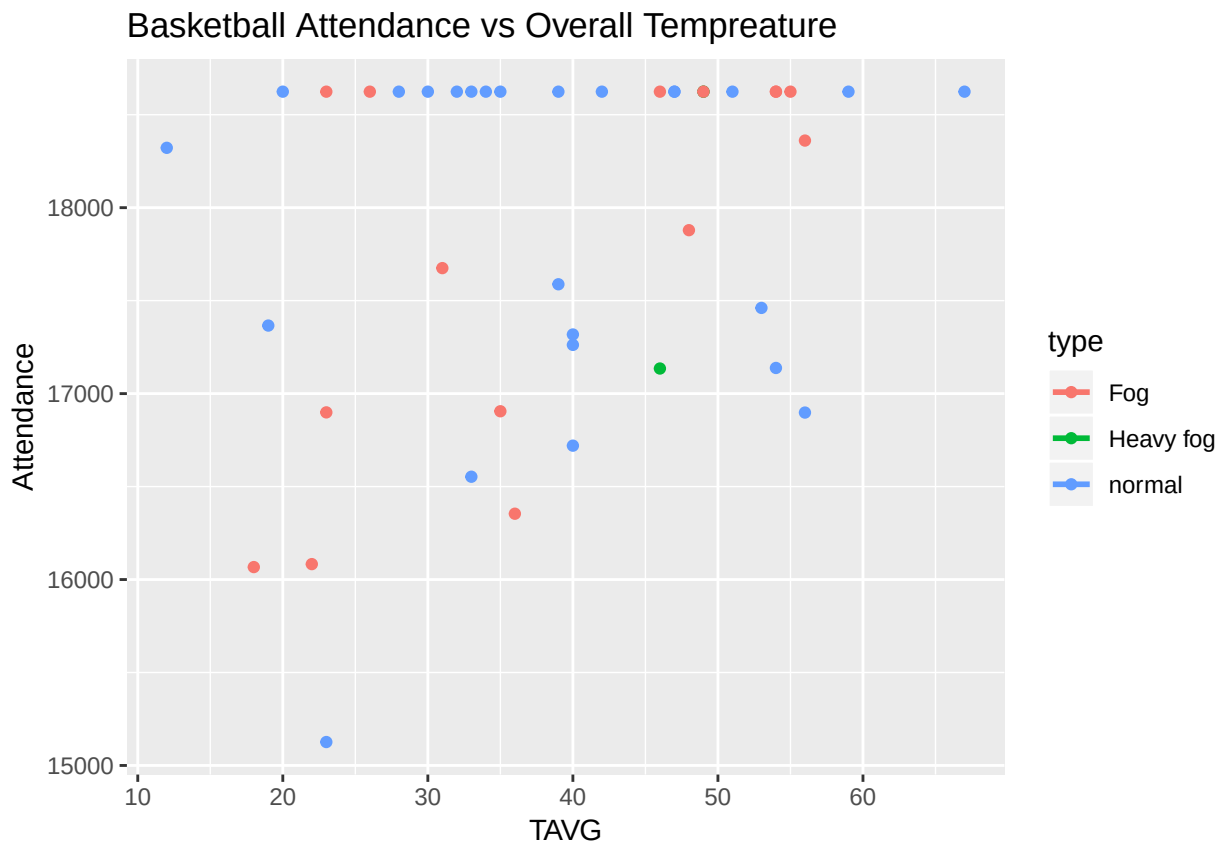

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



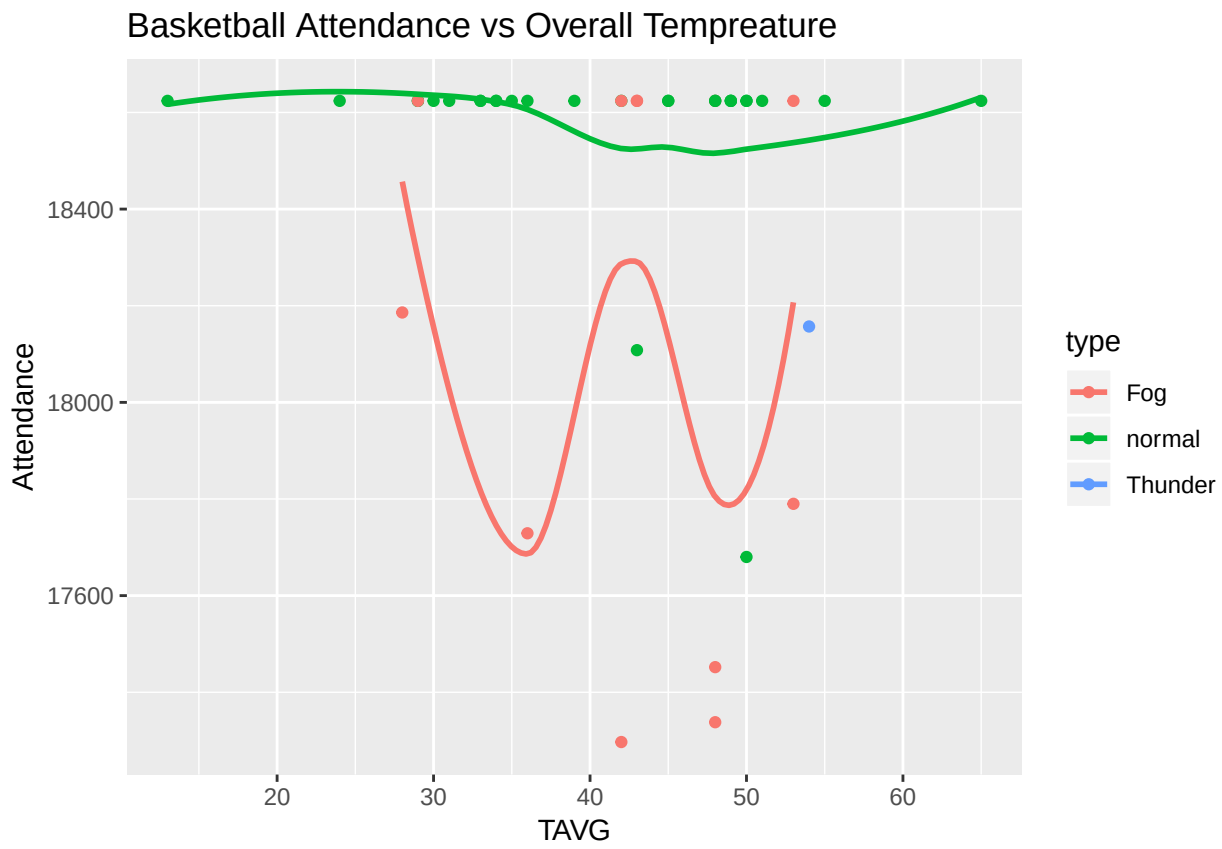

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



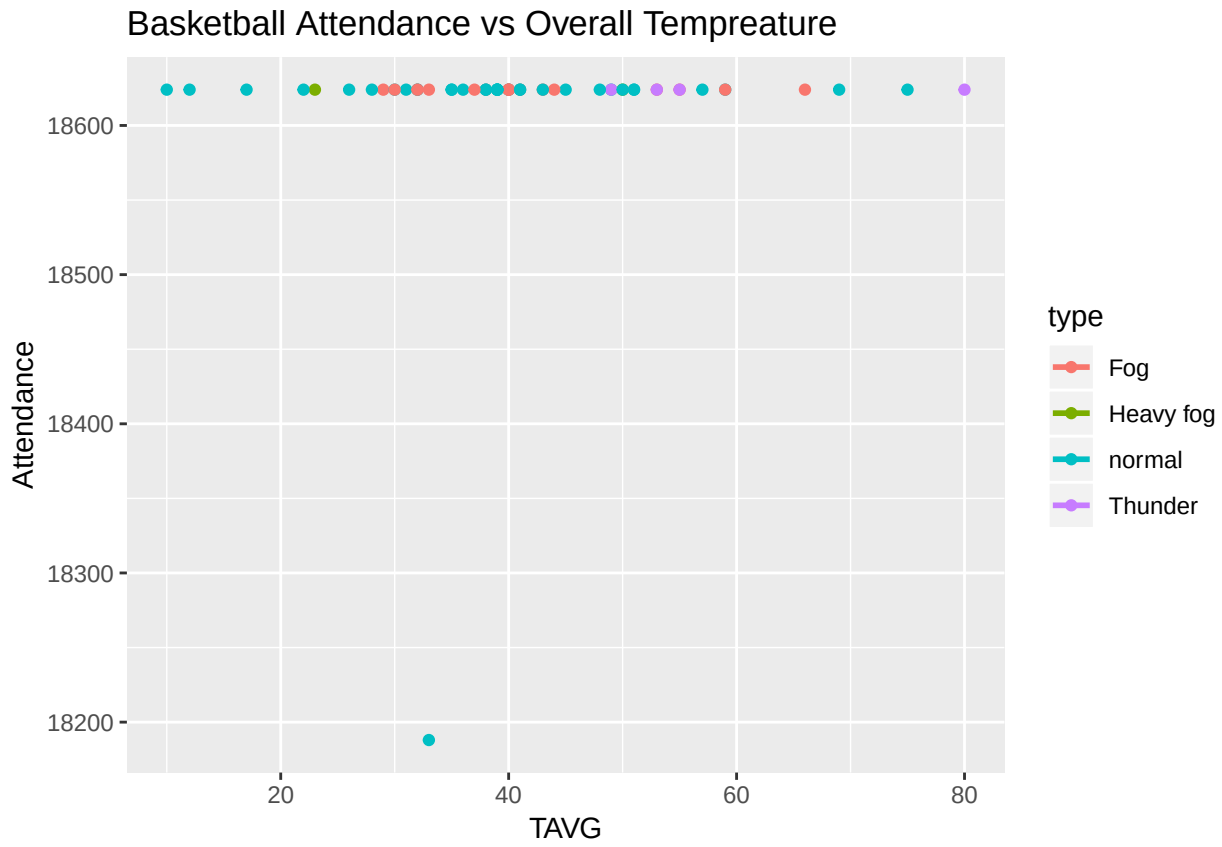

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




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




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



Distribution Table:

Weather Data & EDA —- Wenjia xie& siwei Hu;

Data Cleaning —- Kaiyu Yan& Si Chen;

Basketball data& EDA —- Wenjia xie& Siwei Hu;

BaseBall data& EDA —- Si Chen& Kaiyu Yan;

Shiny Application —- Si Chen & Kaiyu Yan;

Report —- Wenjia xie& Siwei Hu;

The contribution Estimate:

Kaiyu Yan —- 25%; Si Chen —- 25%; Wenjia Xie —- 25%; Siwei Hu —-25%;