# 615Midtermmmmmmm

*Kaiyu Yan,Si Chen, Wenjia Xie, Siwei Hu*

*October 17, 2018*

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

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")
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
##Run through all types to get the weather of a certain day, add that to the "type" column
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"
}
```

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

```r
####### Baseball Data Only #######
Baseball <- read.csv("Baseball.csv")
Clean_1 <- select(Baseball, Date, Tm:Opp, Attendance,Year)
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)))
```

```
## Warning in rbind(c("", "Apr", "5"), c("", "Apr", "7"), c("", "Apr", "8"), :
## number of columns of result is not a multiple of vector length (arg 1)
```

```r
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")

##### 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))))

}

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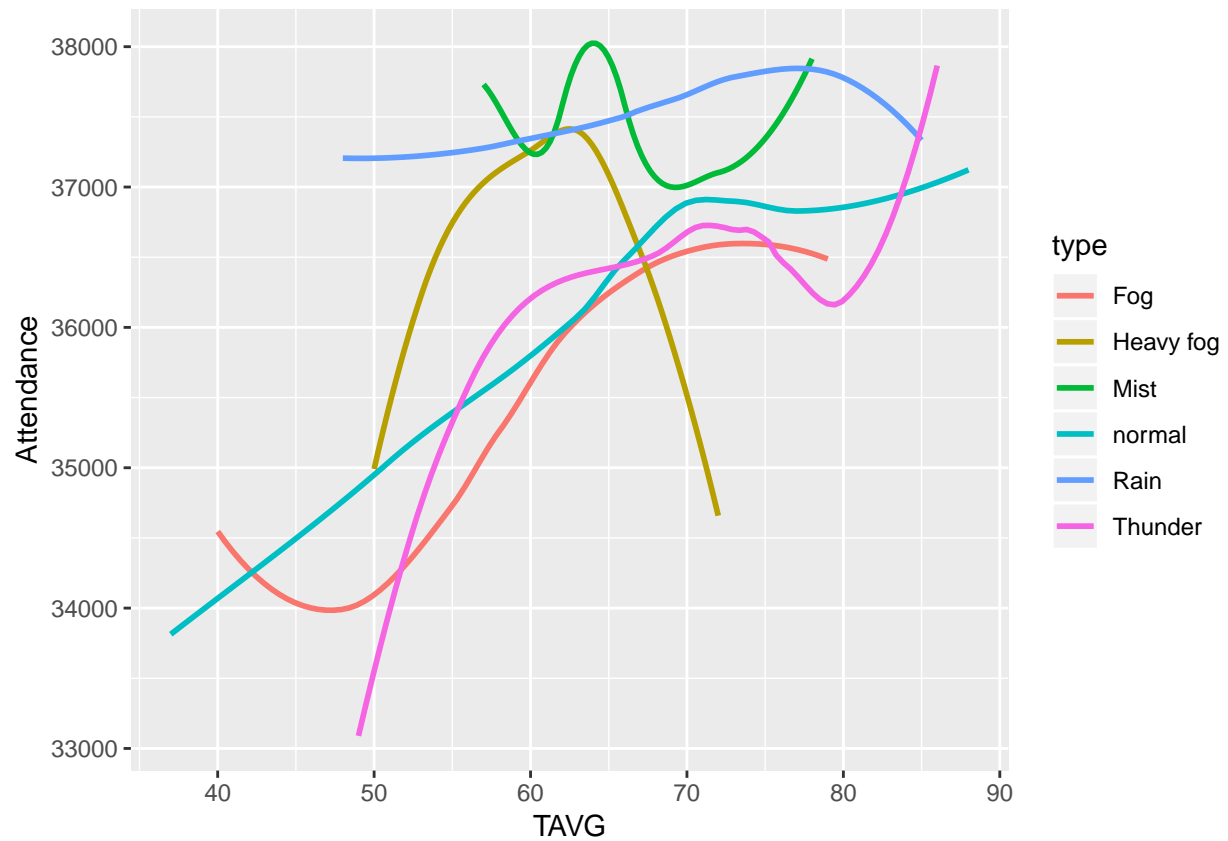
```
ggplot(data = Redsox_All) +
  geom_smooth(mapping = aes( x = TAVG, y = Attendance, color = type), se = F)
```

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

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : Chernobyl! trL>n 6

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : Chernobyl! trL>n 6

## Warning in sqrt(sum.squares/one.delta): NaNs produced

```
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()
    print(p)
    }
```