Assignment: MA 615 Mid-term Project

Xinyi Wang, Qixuan Zhang, Shiyu Zhang, Hao Qin

MIDTERM PROJECT

What is known to all is that Boston is a diverse city with kinds of universities and students from all over the world gathered here to study together. In addition to learning, Boston is also a sports capital, with the baseball team called Red Sox and NBA basketball team called Celtics which have fascinated people for a long time. To find more detail about the baseball and the basketball in Boston, our project aims at exploring the relationship between weather conditions and attendance. Our focus is on attendance at Red Sox and Celtics games.

Our goal for this project is to explore the relationship between weather conditions and attendance at sporting events. The focus of the project is the attendance at Red Sox and Celtics games, and build a dataset that covers the 2012 seasons through the 2017 seasons.

Collecting, Organizing and Exploring the Data

For Red Sox

In order to collect the attendance data for Red Sox, our first step is to acquire and collect relevant data from website for analysis. Collecting useful and good quality data is critical for the project in order to achieve our mission because data analysis typically drives decision-making processes and efficiency optimizations. Our main resource for building Red Sox's dataset was from *the official Red Sox Site, and Baseball Reference*.

We started from exploring all the relevant website about baseball and choose *Red Sox Attendance 2017* website as our starting point.

Once we got the attendance data directly from the URL from the official website in R. In R studio, we first change the data frame by separating the baseball stats URL and adding years between the two links to make the baseball dataset more readable.

Then we used “For Loop” function in R to read through each URL to get "Attendance", "home" and "Date" categories to make the dataset more formalized. Next, we took a further step to clean and modify baseball dataset such as change the "Date" format, change column names, etc.) to make the datasets more visible to and usable for anyone who’s interested.

```
#Baseball Data Scrape from Web
```{r}
##Separate Baseball stats URL
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 = "")

##Run through each URL to get "Attendence","home",and "Date"
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 and modify baseball dataset(change "Date" format,change column names, etc.)
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))
colnames(baseball)[2] = "DATE"
colnames(baseball)[3] = "NA"

```
```

For weather data related to Red Sox's events, except the NOAA weather buoy Data, we also checked The *National Center for Environmental Information* and some other website and choose the one which we think the most suitable for the analysis.

Finally, we chose data from *noaa.gov* and pulled out of its original Excel format, tidied, and uploaded. What's more, we also checked how tables were named and how the excel was filed, etc. Then, we imported the csv file directly into R.

The next step is to clean and format the data. We first cleaned and modified the weather dataset by changing "Date" format, removing columns with all N/A, etc.

Next, we built two vector contains all weather type code and corresponding actual type and added new column named "type" and set them to all N/As. Then, we run through all types to get the weather of a certain day and added them to the "type" column by using "For Loop" function in R. Final step is to select the only columns that matter to this project. Now the datasets are ready for analysis.

```
#Weather Scrape from Web
```{r}
##Donloaded "weather.csv" from noaa.gov,read csv file
weather = read.csv("weather.csv",header=TRUE)

##Clean and modify weather dataset(change "Date" format,remove columns with all NA, etc.)
weather$DATE = as.Date(weather$DATE,format="%Y-%m-%d")
weather = weather[,colSums(is.na(weather)) < nrow(weather)]

##Build two vector contains all weather type code and corresponding actual type(information are from noaa.gov).
Add new column named "type" and set them to all NA.
typelist = c("Fog","Heavy Fog","Thunder","Ice
Pellets","Hail","Glaze","Smoke","Blowing","Mist","Drizzle","Freezing Drizzle","Rain","Freezing Rain","Snow",
         "Unknow Source of Preipitation","Ice Fog")
type_code = c("WT01","WT02","WT03","WT04","WT05","WT06","WT08","WT09","WT13",
              "WT14","WT15","WT16","WT17","WT18","WT19","WT22")
weather$type<-NA

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

weather[is.null(weather)] <- 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)[1]) {
  t<-0
  for (n in 1:length(typelist)) {
    if (is.null(weather[m,typelist[n]])) {
      weather[m,typelist[n]] = NA
    }
    if (!is.na(weather[m,typelist[n]])) {
      weather[m,"type"] =  typelist[n]
      t<-t+1
    }
  }
  if(t==0)
    weather[m,"type"] = "normal"
}

##Select only columns that matters to this project
weather = weather[,c("DATE","TMAX","TMIN","type")]
x = cbind(weather$TMAX,weather$TMIN)
weather$tavg = apply(x,1,mean)
```
```

For Celtics

For the basketball, the first step is to find the dataset from the website. In ESPN, which is http://www.espn.com/nba/team/schedule/_/name/bos/season/ . There are so many variance data about the different teams and the different seasons, so in order to make our project more accuracy, we pick the time from 2012 to 2018, pick the game happened in Boston, and then we rebuild the URL into a new one. The second step is to find the weather dataset which can influence the basketball, so we search the website NOAA, which is https://www.noaa.gov/ . The NOAA's National Weather Service is to build a Weather-Ready Nation by providing better information for better decisions to save live and livelihoods, "Each year, the United States averages some 10,000 thunderstorms, 5,000 floods, 1,300 tornadoes and 2 Atlantic hurricanes, as well as widespread droughts and wildfires. Weather, water and climate events, cause an average of approximately 650 deaths and $15 billion in damage per year and are responsible for some 90 percent of all presidentially-declared disasters", from the homepage states that the weather is always a significant event in our life, this is also the reason why we pick the weather as our variable to compare with the attendance in basketball, so we download the data about the Boston from the NOAA, you can just have a look at the screenshots about the part of the excel in the weather.

B	C	D	E	F	G	H	I	J	K	L	M
NAME	LATITUDE	LONGITUD	ELEVATION	DATE	AWND	AWND_AT	PRCP	PRCP_ATT	SNOW	TMAX	TMIN
BOSTON, M	42.3606	-71.0097	3.7	2012/1/1	9.17	,,W	0.01	,,X,2400	0	52	39
BOSTON, M	42.3606	-71.0097	3.7	2012/1/2	13.87	,,W	0.01	,,X,2400	0	50	34
BOSTON, M	42.3606	-71.0097	3.7	2012/1/3	14.76	,,W	0	,,X,2400	0	35	14
BOSTON, M	42.3606	-71.0097	3.7	2012/1/4	11.86	,,W	0	,,X,2400	0	28	10
BOSTON, M	42.3606	-71.0097	3.7	2012/1/5	11.41	,,W	0	,,X,2400	0	39	25
BOSTON, M	42.3606	-71.0097	3.7	2012/1/6	5.82	,,W	0	T,,X,2400	0	48	28
BOSTON, M	42.3606	-71.0097	3.7	2012/1/7	4.92	,,W	0	,,X,2400	0	60	30
BOSTON, M	42.3606	-71.0097	3.7	2012/1/8	11.63	,,W	0	,,X,2400	0	45	30
BOSTON, M	42.3606	-71.0097	3.7	2012/1/9	7.16	,,W	0	,,X,2400	0	40	25
BOSTON, M	42.3606	-71.0097	3.7	########	10.07	,,W	0.02	,,X,2400	0.5	47	30
BOSTON, M	42.3606	-71.0097	3.7	########	7.61	,,W	0	,,X,2400	0	38	30
BOSTON, M	42.3606	-71.0097	3.7	########	19.46	,,W	0.92	,,X,2400	0	42	33
BOSTON, M	42.3606	-71.0097	3.7	########	18.34	,,W	0.05	,,X,2400	0	54	31

(Figure 1)

Then, after collecting the basketball data and the weather data, we should plug all the information into R, let R tell us what's going on about the relationship between attendance and the weather. Since all the dataset we collected from the web is raw data, we must acquire clean and organize the data. For the basketball dataset, the first step, we deleted the cancelled game and unwanted home variable, and also there are many blank, called "home", we should delete that too. The second step is to combine all years together, read the attendance variable, and use the code "join" to combine the worked basketball and the weather, for more details about the variable and the characters, you can see the Figure 2, there are nine different variables, the TMAX means the highest temperature, the TMIN means the lowest temperaure, and the tavg means the avearge of temperature

```
names(basketball_weather)

## [1] "gameID"     "DATE"       "home"       "YYYY"       "Attendence"
## [6] "TMAX"       "TMIN"       "type"       "tavg"
summary(basketball_weather$type)

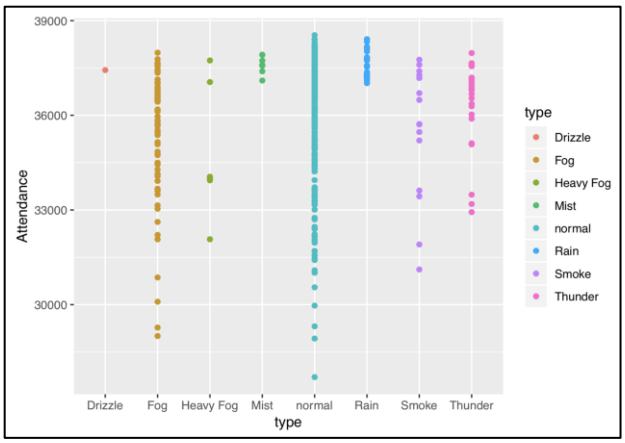
##    Length     Class      Mode
##       258 character character
```

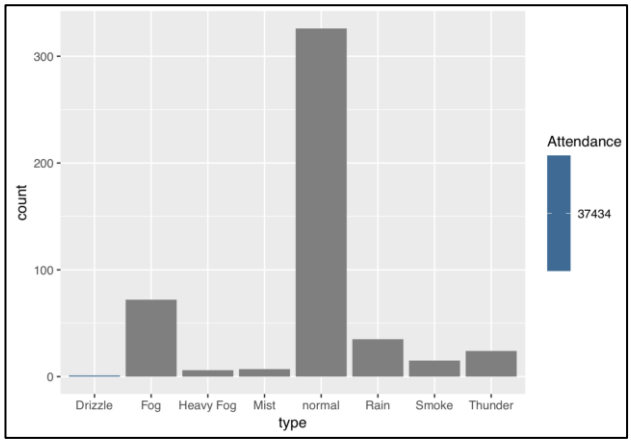
(Figure 2)

Data Visualization

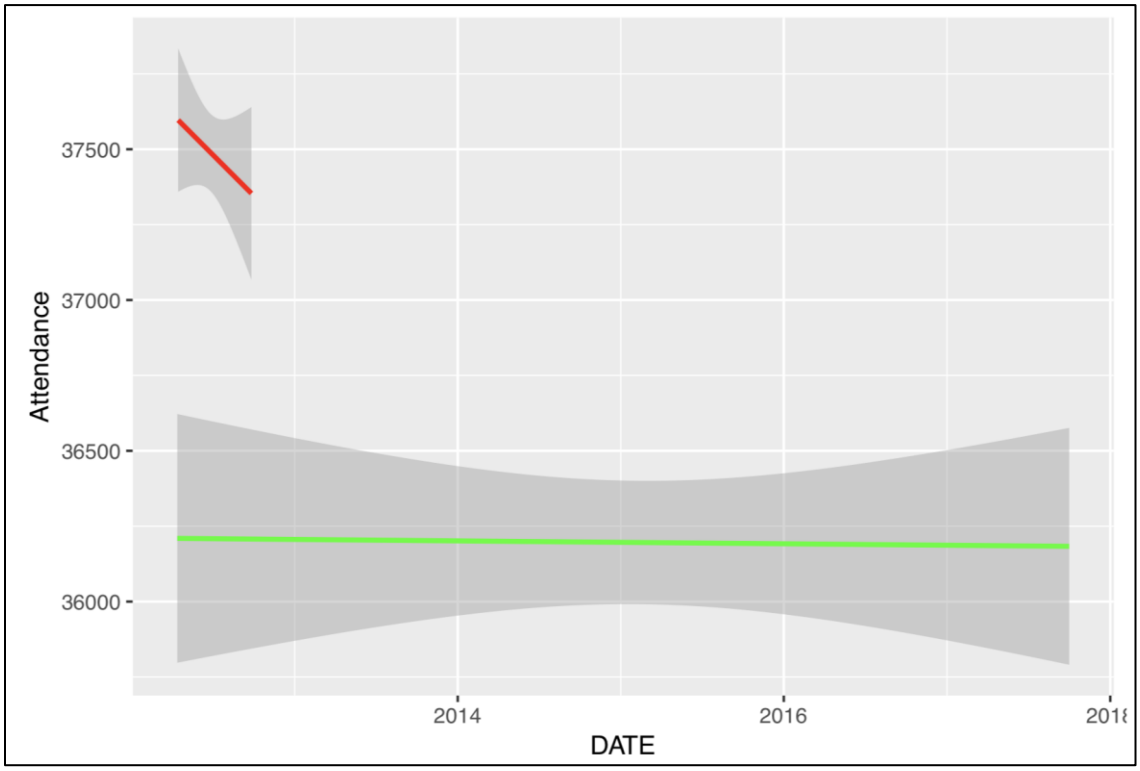
For Red Sox

To explore the relationship between the attendance and the weather, we built several plots to visualize the difference in attendance with different weather by using "ggplot" in R.

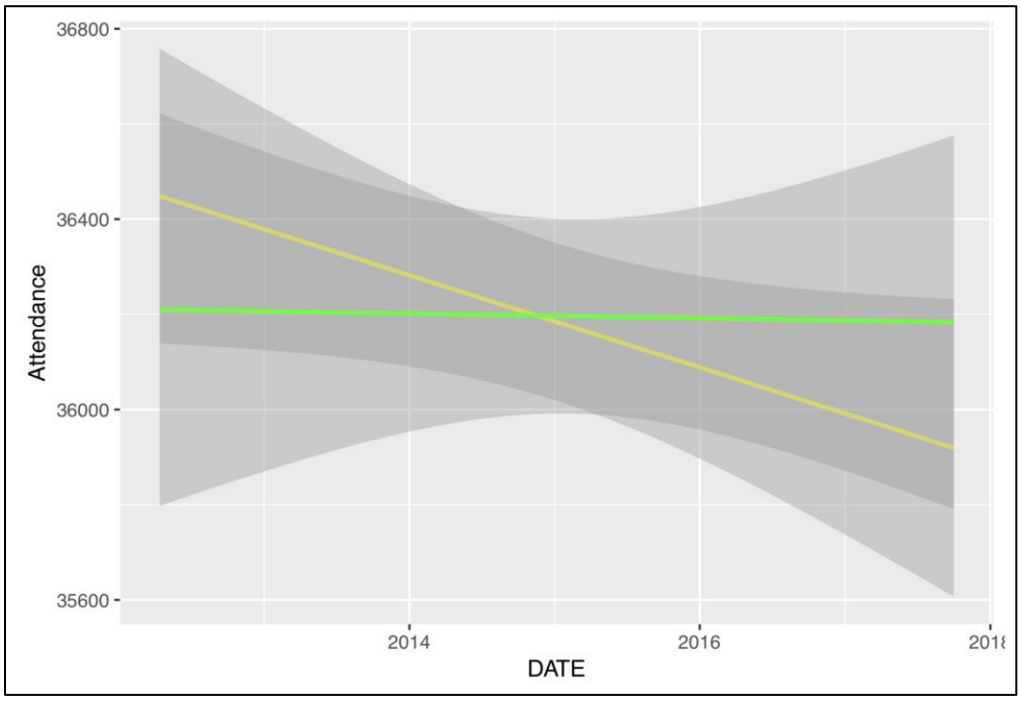




In the graph on the left above, different colors represent different weather conditions such as green for fog and pink for Thunder. From the right-side histogram, we can see that most of the points lies in normal weather and Fog weather. However, from the left side graph we can find that even sometimes the weather is rain, smoke or even thunder, the attendance still can be very high. But in general, good weather has high attendance.

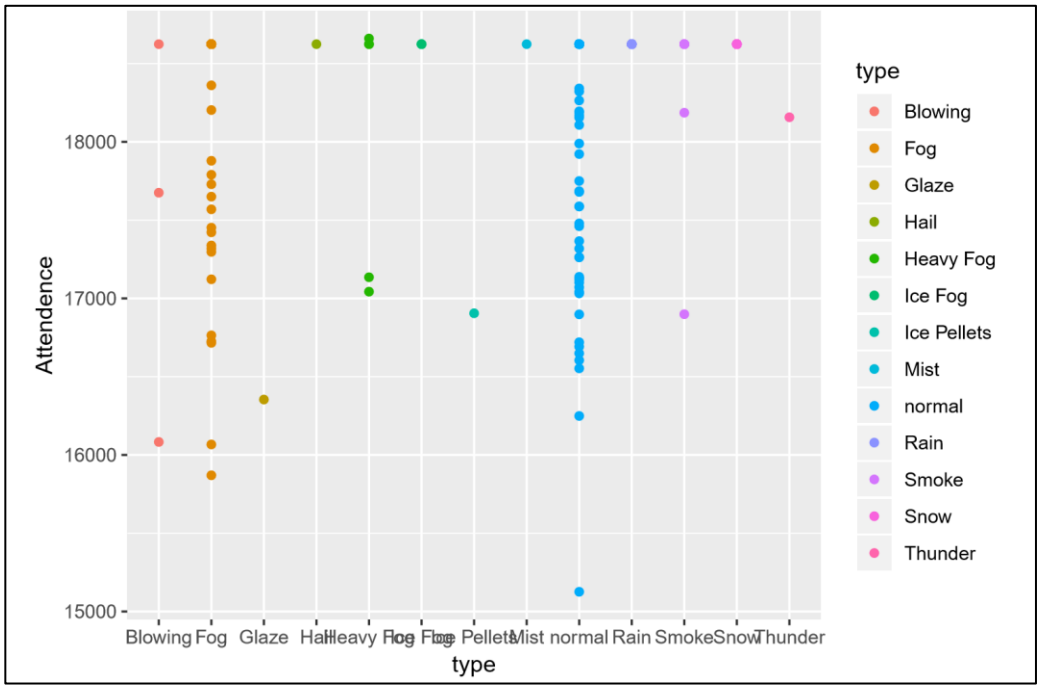


To compare the difference between normal weather and rain weather condition, the red line stands for rains and green line stands for normal weather. We found that, the attendance under the normal weather condition remains stable, but for rainy days, the attendance has a negative trend from 2012 and 2013. This situation might because baseball stadium is outdoors and have a strong relevance to weather, so the weather has impact on the attendance for Red Sox games.



We took a further step to explore the relationship between fog and normal weather. The yellow line stands for fog weather condition, the green line stands for normal weather. From the graph, we found that the attendance for normal is stable, however the attendance under fog condition went down from 2012 to 2018. We think that it makes sense because the fog weather has a bad impact for audience to watch the game.

For Celtics

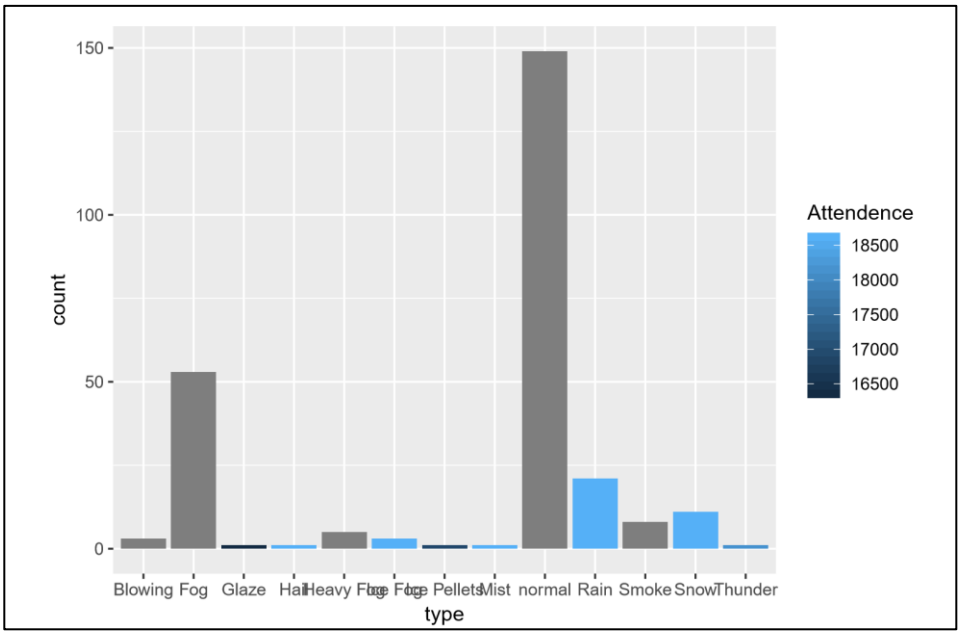


(Figure 3)

From the plot (Figure 3), you can figure out that different colors represent different weather conditions such as green for fog and pink for Thunder. Most of the points located in the normal type and the fog type, I think most organizers like to say that the games are held in good weather, but there is one point located in the lowest part in normal, I think this is an error point, just ignore that. Overall, under any weather conditions, there is a change in the level of attendance, but overall, the attendance is relatively high, probably because basketball is indoor, so the weather is not too big.

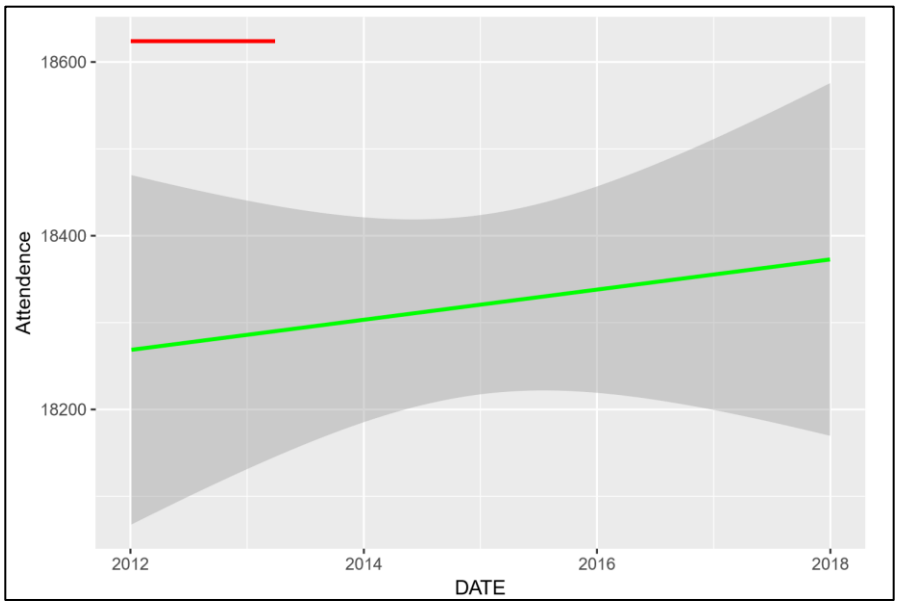
Sometimes the dot plot is not easy to see what the difference in attendance is under different weather conditions, so to express this relationship better, we decided to create the histogram plot to show what it is (Figure 4). From the results of the histogram chart, we can conclude that the attendance rate in the first place is normal weather, the second

place is foggy weather, the third place and the fourth place is rain and snow, other weather accounts for a relatively small proportion



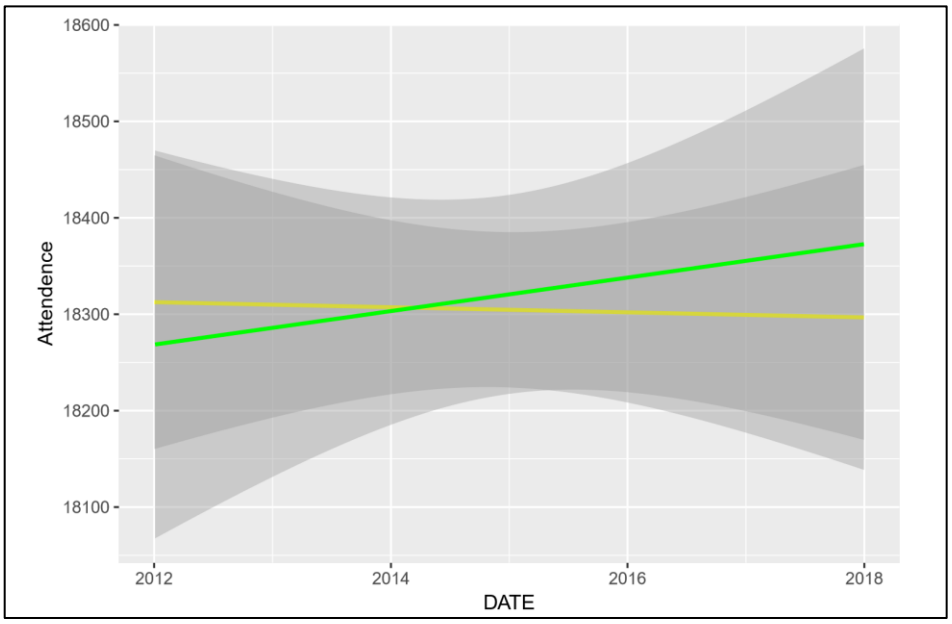
(Figure 4)

In order to compare the different weather, so first I pick the normal and the rain, I use the green line to show the normal weather, and the red line to show the rain weather, you can see from the figure 5, with the time goes by, the attendee under the normal is increasing, but for the red line, this is only appeared in 2012 and 2013. If I put the snow weather as the blue line, the red and blue lines will overlap.



(Figure 5)

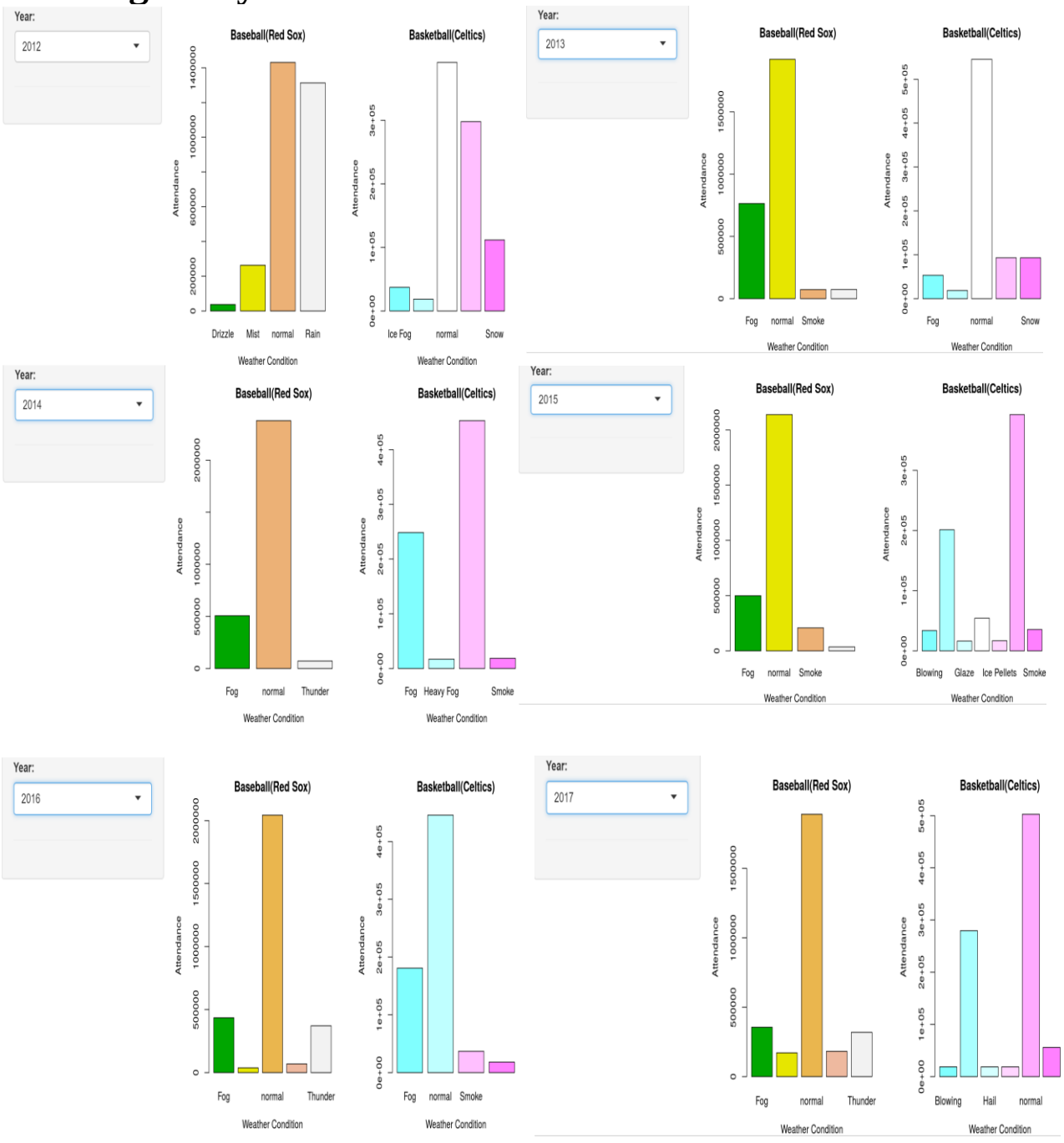
For the last part, I would like to compare the top 2 weather, those are rain the foggy, you can see from the Figure 6



(Figure 6)

I represent the green line for rain and the yellow line for the fog, so from the outcome, we can get that with the time goes by, there is no difference in attendance under the fog weather, but for the normal, it is increasing.

Building Shiny APP:



Those plots are from our shiny website, which shows the attendance according to the different weather condition for Red Sox and Celtics. From the year 2012 to 2017, when the weather is normal, it occupied the largest proportion of attendance, it is easy to see the different attendance under the Shiny APP.

In conclusion, from the data we collect and from the plot we create, we figure that although the weather can have a small effect on attendance, it is not a decisive factor.