Midterm Project

Longhao Chen, Angela Zhai,Yudi Mao, Tianying Xu

2018.10.18

I. Data

For baseball attendance data, we download it from http://www.espn.com/mlb/team/schedule/_/name/bos.

For weather data, we download it from NOAA website, the url is https://www.ncdc.noaa.gov/cdo-web/datasets. The data includes a lot of index like temperature, windspeed,snowfall, and it starts at 2012 and ends at 2018.

For basketball attendance data, we get the data from ESPN website using web crawler, with the url http://www.espn.com/nba/team/schedule/_/name/bos/season/2018. We use rvest package to craw the attendance data through 2012-2013 season to 2017-2018 season.

II. Analysis

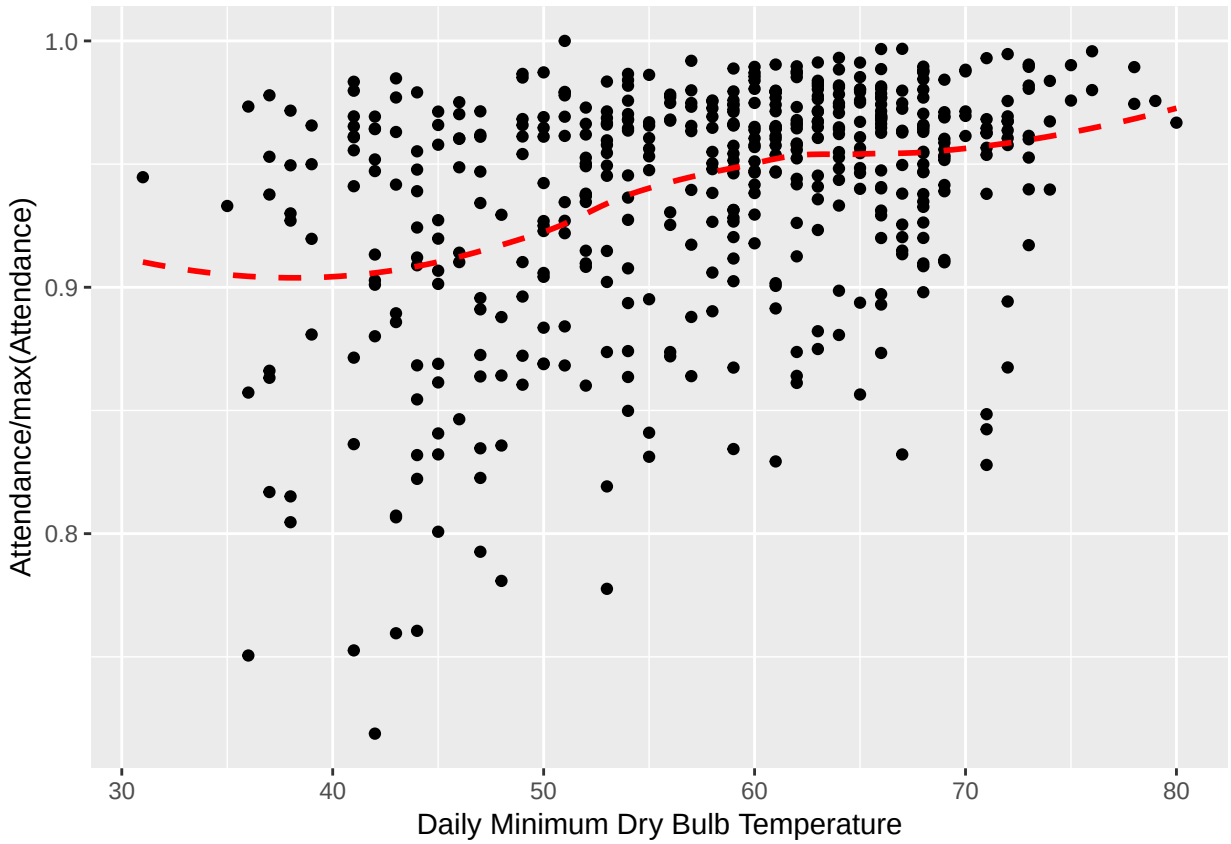
1. Baseball

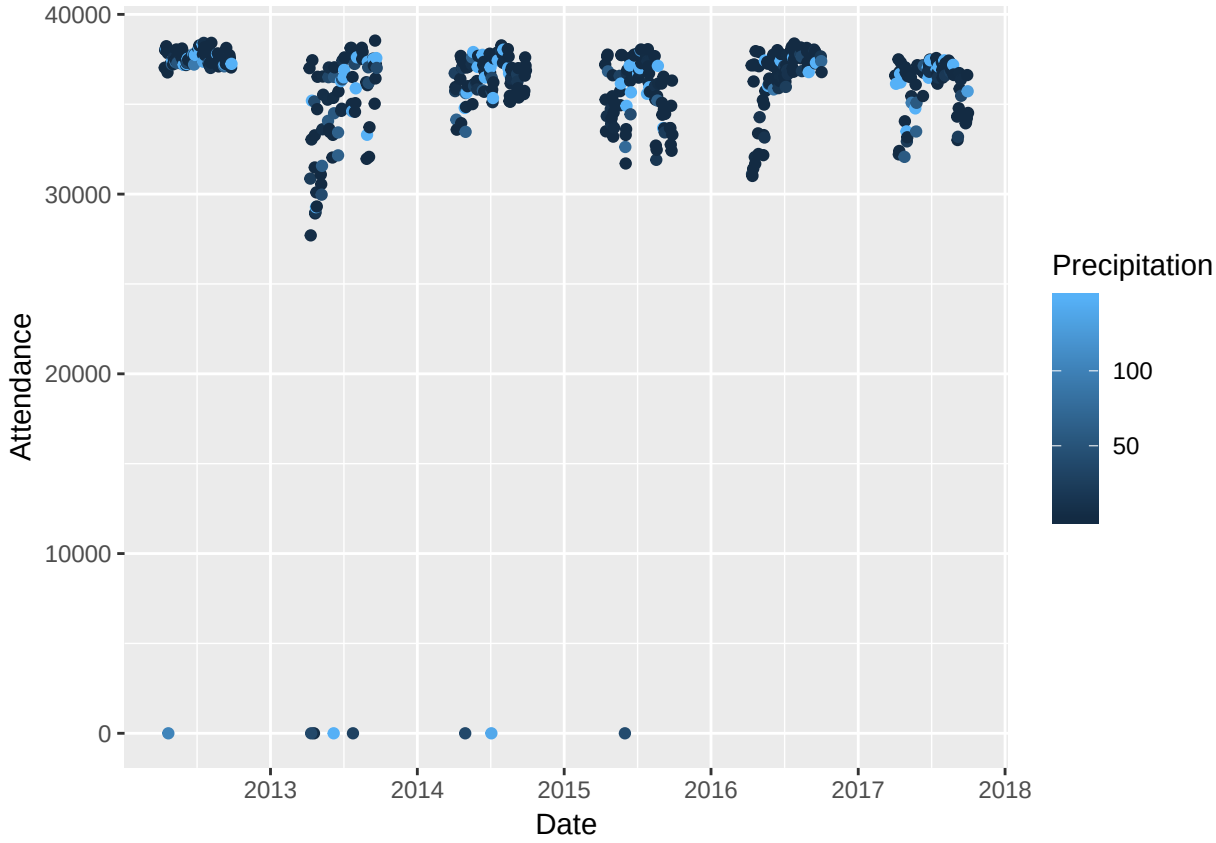
a. data

Wrangled in the I.DATA part. We transform date in both redsox attendance and weather using as.Date, and join these two parts with the DATE as the key.

b. EDA

We can roughly conclude that low temperature has relationship with low attendance, while high temperature vice versa.


Attendance/max(Attendance)
1.0
0.9
0.8
30
40
50
60
70
80
Daily Minimum Dry Bulb Temperature


Attendance
40000
30000
20000
10000
0
2013
2014
2015
2016
2017
2018
Date
Precipitation
100
50

2. Basketball

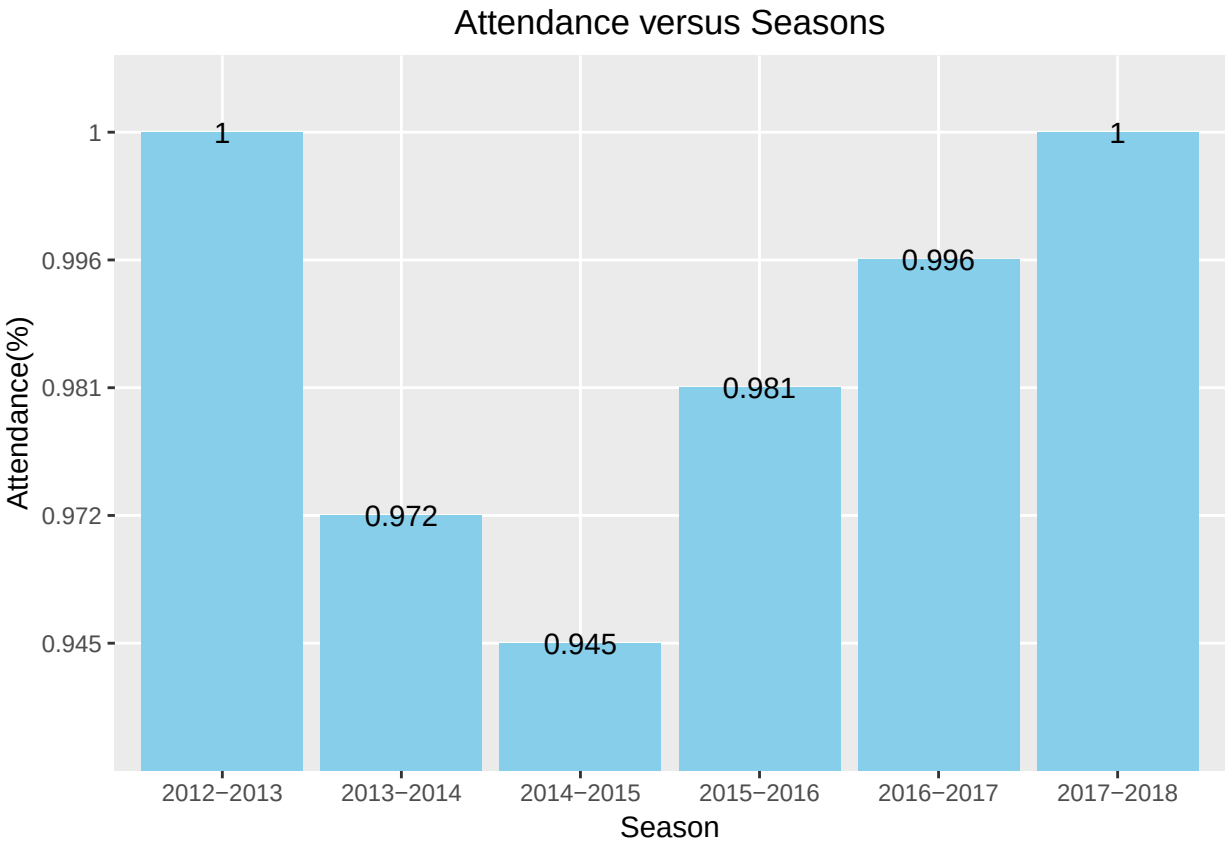
a. data

We need to join the weather and basketball data together by date so that we could do EDA. We first translate the date of basketball and weather to the same format and join them, then clean data. (There are "T" in the value of index that means the number is too small to measure, thus we translate these values to be 0.)

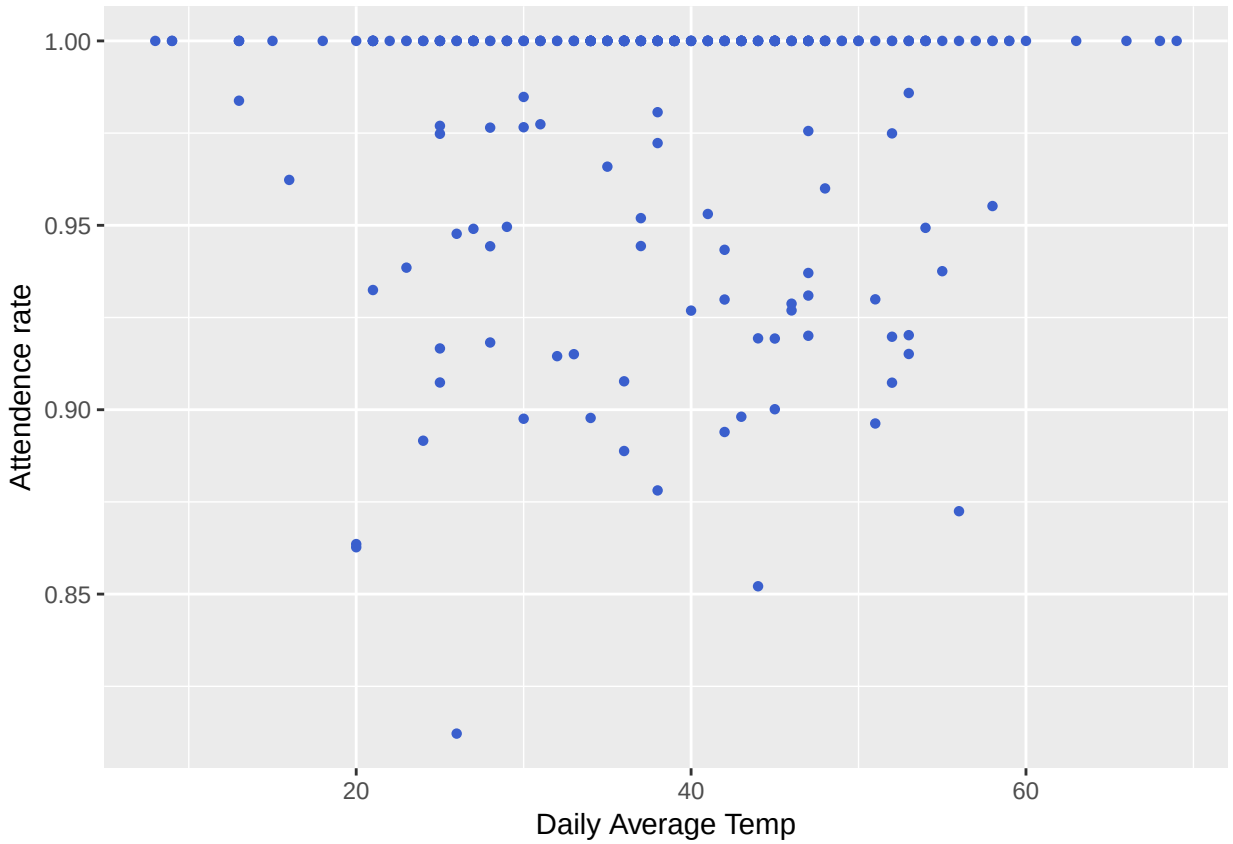
b. EDA

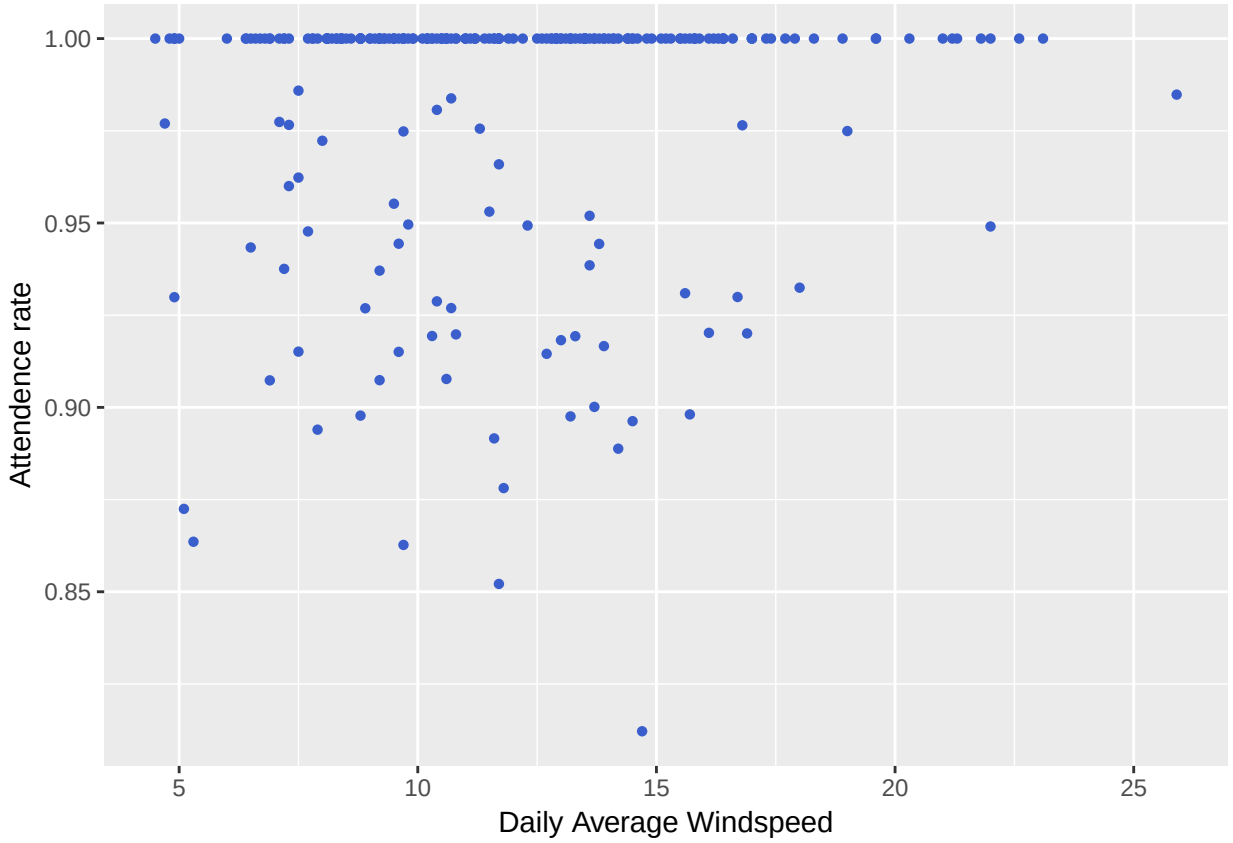
For weather data, we choose daily average normal temperature, daily average wind speed, daily average snow fall.

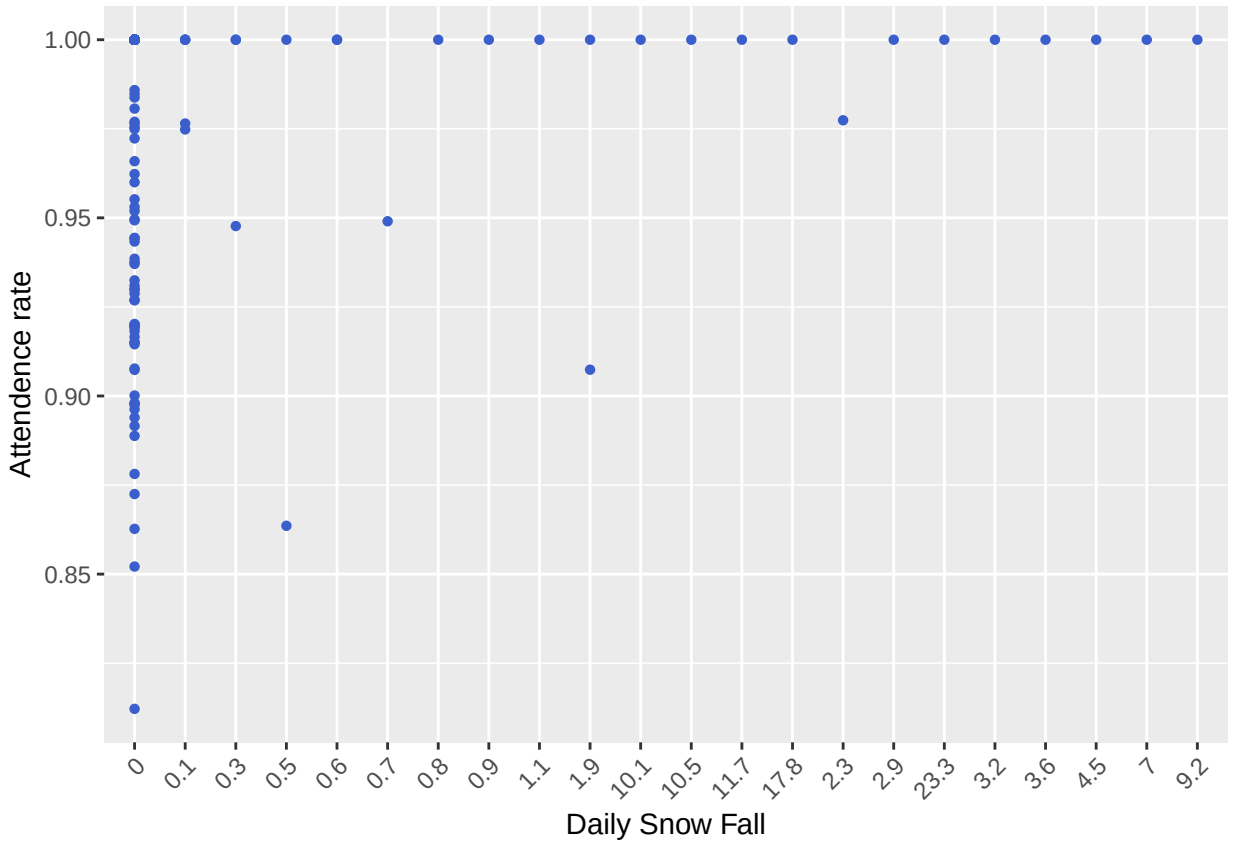
First we plot average attendance through each season.



Then we plot attendance versus every weather variable to see the association between attendance and weather.

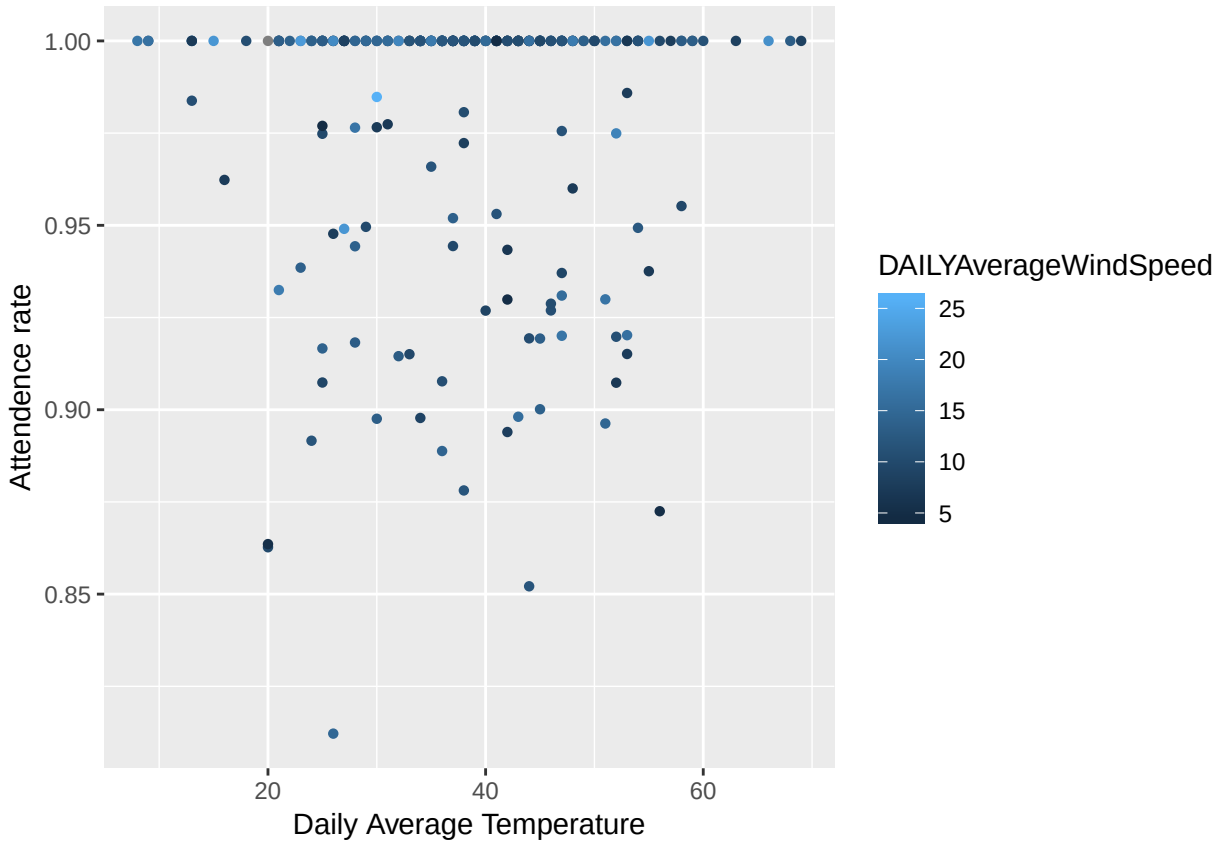

1.00
0.95
0.90
0.85
Attendence rate
20
40
60
Daily Average Temp

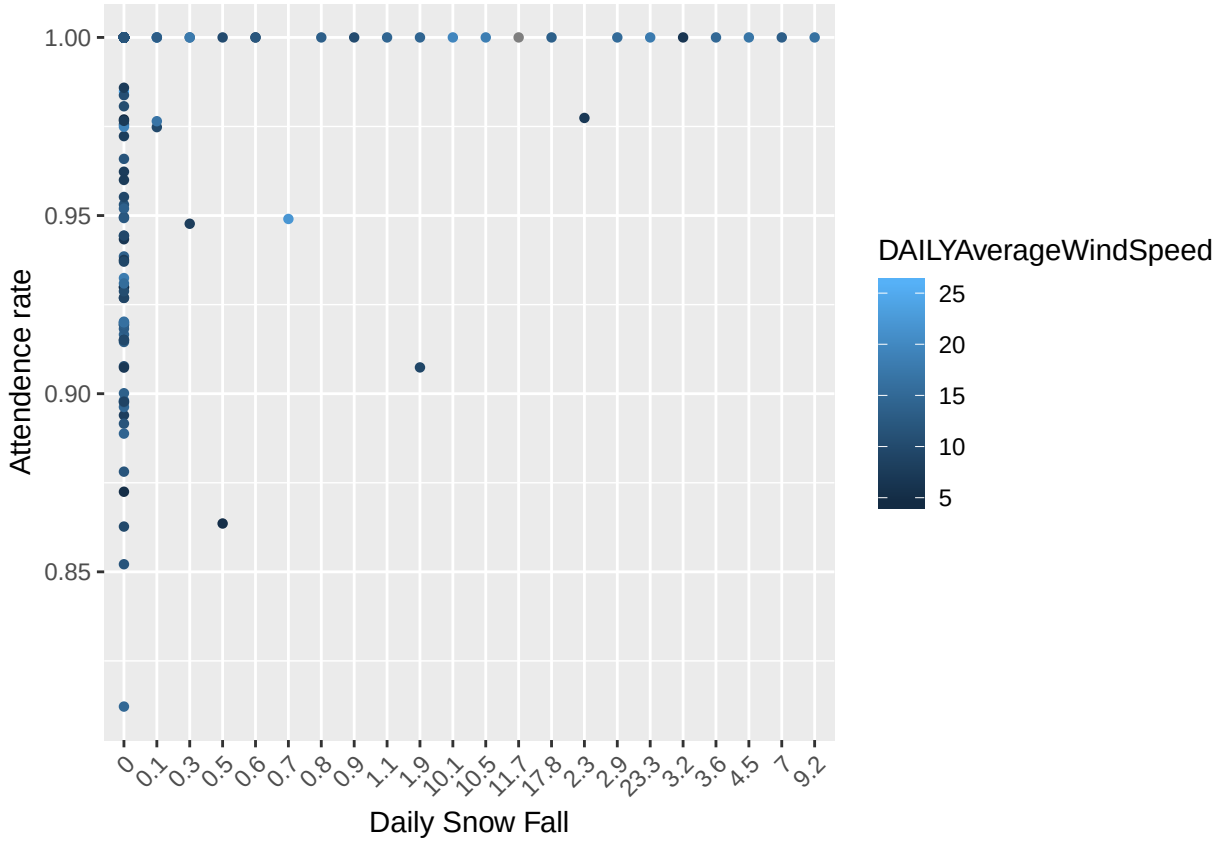

1.00
0.95
0.90
0.85
Attendence rate
5
10
15
20
25
Daily Average Windspeed

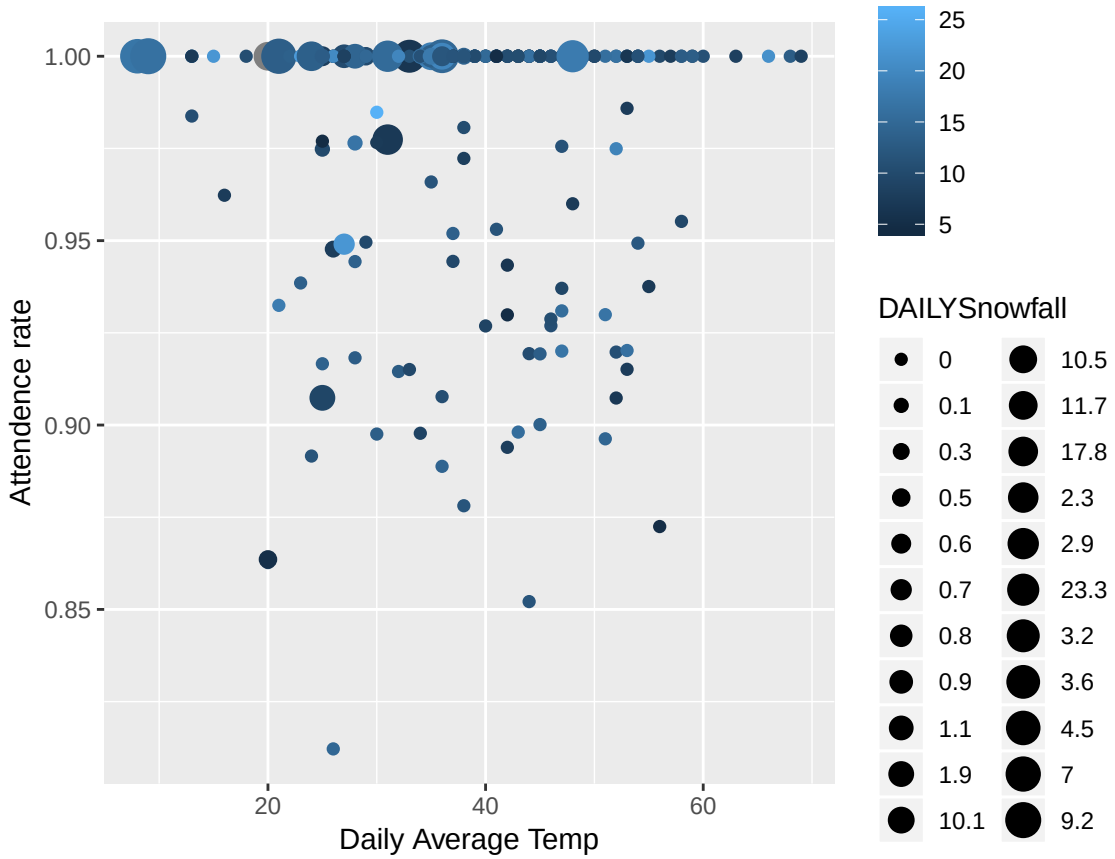


From the plots above, we could see that whatever the weather situation is, the attendance could still be high or low, there seems no correlation between weather situation and basketball attendance.

Finally we try to plot attendance versus two variables to see if there is interaction effects.


1.00
0.95
0.90
0.85
Attendence rate
20
40
60
Daily Average Temperature
DAILYAverageWindSpeed
25
20
15
10
5


1.00
0.95
0.90
0.85
Attendence rate
0
0.1
0.3
0.5
0.6
0.7
0.8
0.9
1.1
1.9
10.1
10.5
11.7
17.8
2.3
2.9
23.3
3.2
3.6
4.5
7
9.2
Daily Snow Fall
DAILYAverageWindSpeed
25
20
15
10
5



These three plots do not show any pattern and association of weather situation and attendance. In the final plot, most points with bigger size and lighter color, which means the snow is heavy and the wind speed is fast, are still correspond to the 100% attendance. This does not make sense. Therefore weather is not highly associated with attendance rate.