615 Midterm Project Report

Dave Anderson, Sky Liu,Tingtui Huang,Xiang Xu

October 19, 2018

Introduction

Lots of research has been done on weather condition as a critical influence to some activities. In our report, we are going to find out if severe weather condition will affect the attendance for sports games. We focus on two types of sports, indoor and outdoor. For indoor sports, we will analyze the attendance of Celtics games and for the outdoor sports we will analyze the attendance of Red Sox games. This report has been done in four parts. First, we will talk about what are our data sources and what we have done to get the data ready to analyze. Second, we conduct EDA on multiple weather conditions and attendance and try to find out the potential relation. Third, we will discuss our outcomes and findings. In the end, we will introduce our interactive data visualization which allows users to play with the data by themselves.

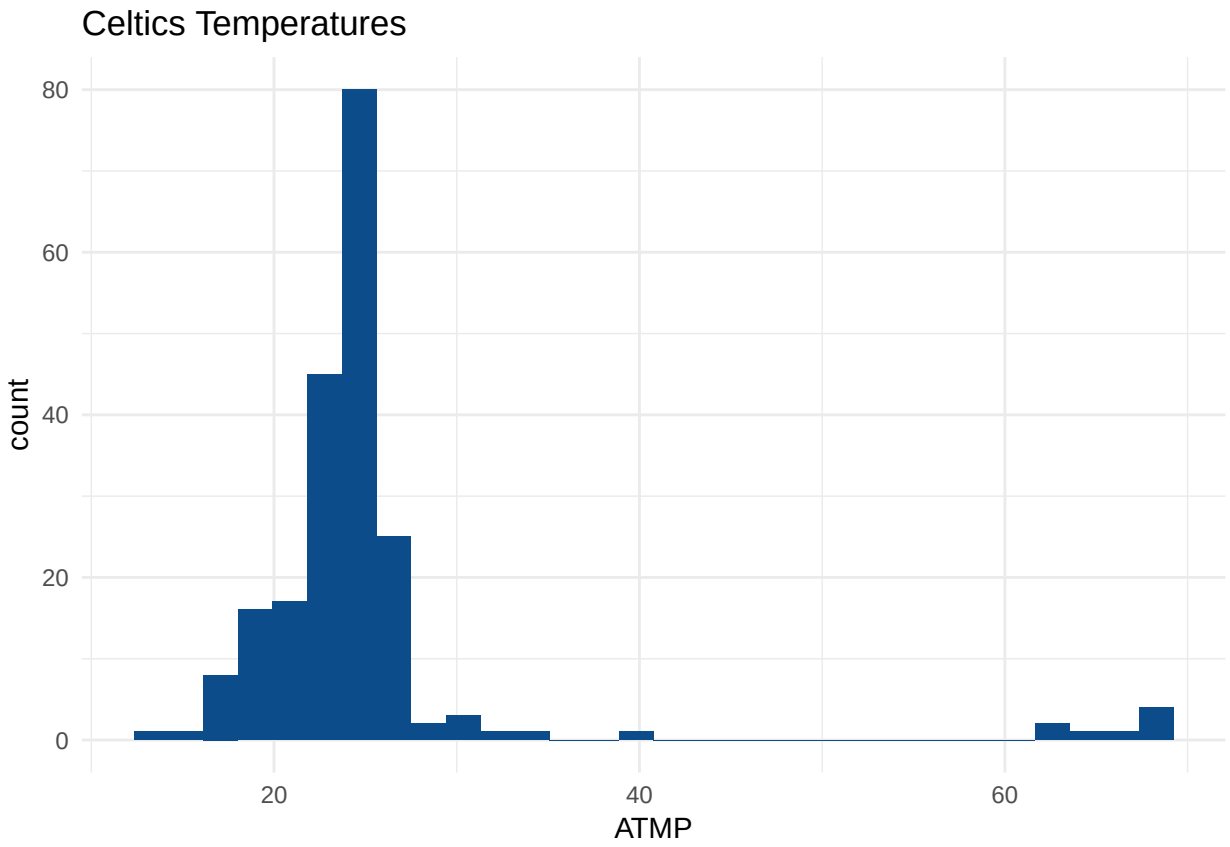
Data Collection and Preparation

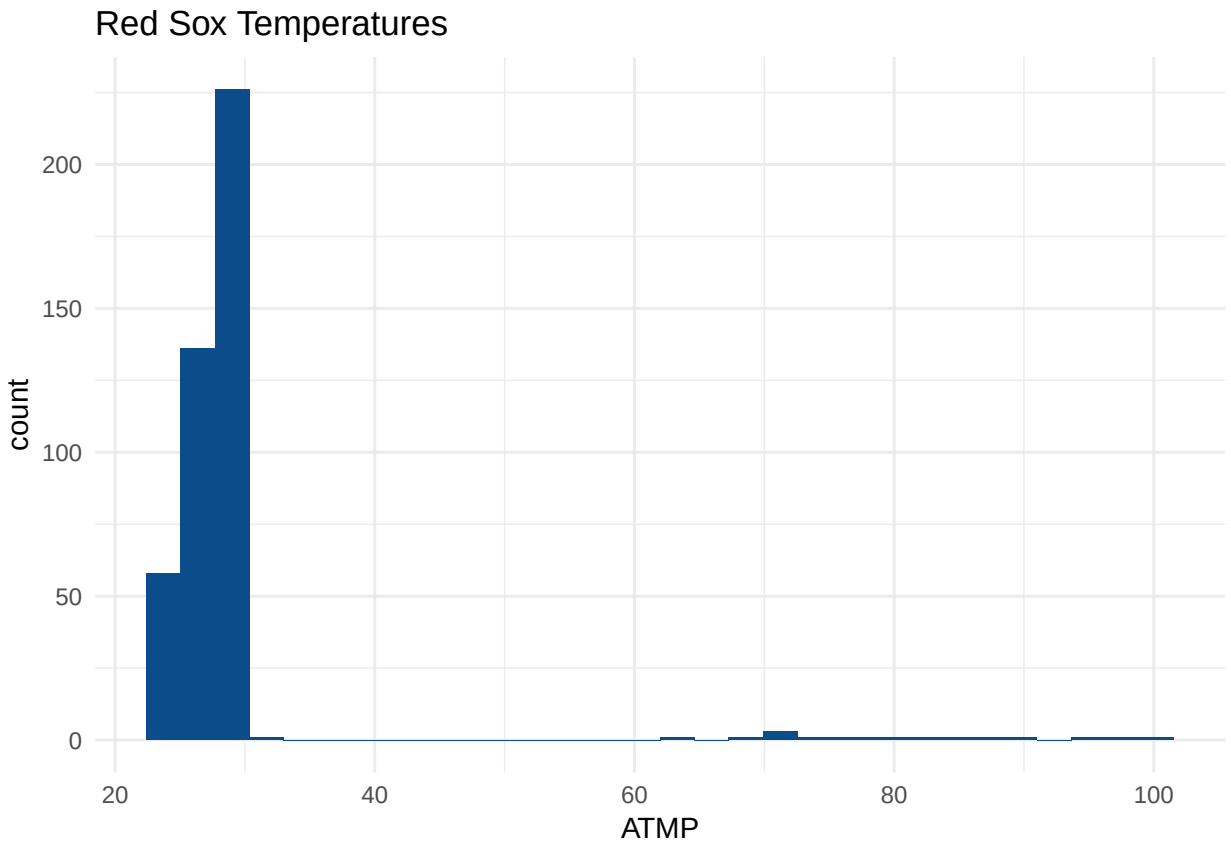
The weather data was sourced from National Data Buoy Center. Since we are not sure about which variables we will use, we simply keep all the variables in the datasets. The air temperature is tested multiples times everyday, thus we took the average of them and assigned each day only one temperature record. Meanwhile, we found the year, month and date were stored in three separate columns, therefore we combined those three columns, formatted them as ???date???, and put them in a new column. The attendance data of Red Sox games were sourced from Baseball-Reference. Totally six datasets were downloaded directly from the website. The six datasets contain the date, day of week, game result, opponent, team rank, team???s overall records in current season and attendance from season 2012 to season 2017. Since we are going to explore the relationship between attendance and weather, we removed some of the unrelated variables such as team rank and overall records. After looked at the dataset in R, we found some format issues then we reformatted most of the variables to proper data type so that R could read them without any issue. The attendance data of Celtics games were sourced from ESPN. We conducted similar data cleaning works for these datasets, for example, we removed some unrelated variables, merged datasets into one table and formatted variables into proper data type. As we have all three datasets ready, we merged weather data with each of the Red Sox and Celtics dataset. We didn???t merge all three into one dataset because the dates of games for each sport are different, and if we merged them together there would be lots of NAs in the dataset.

Exploratory Data Analysis

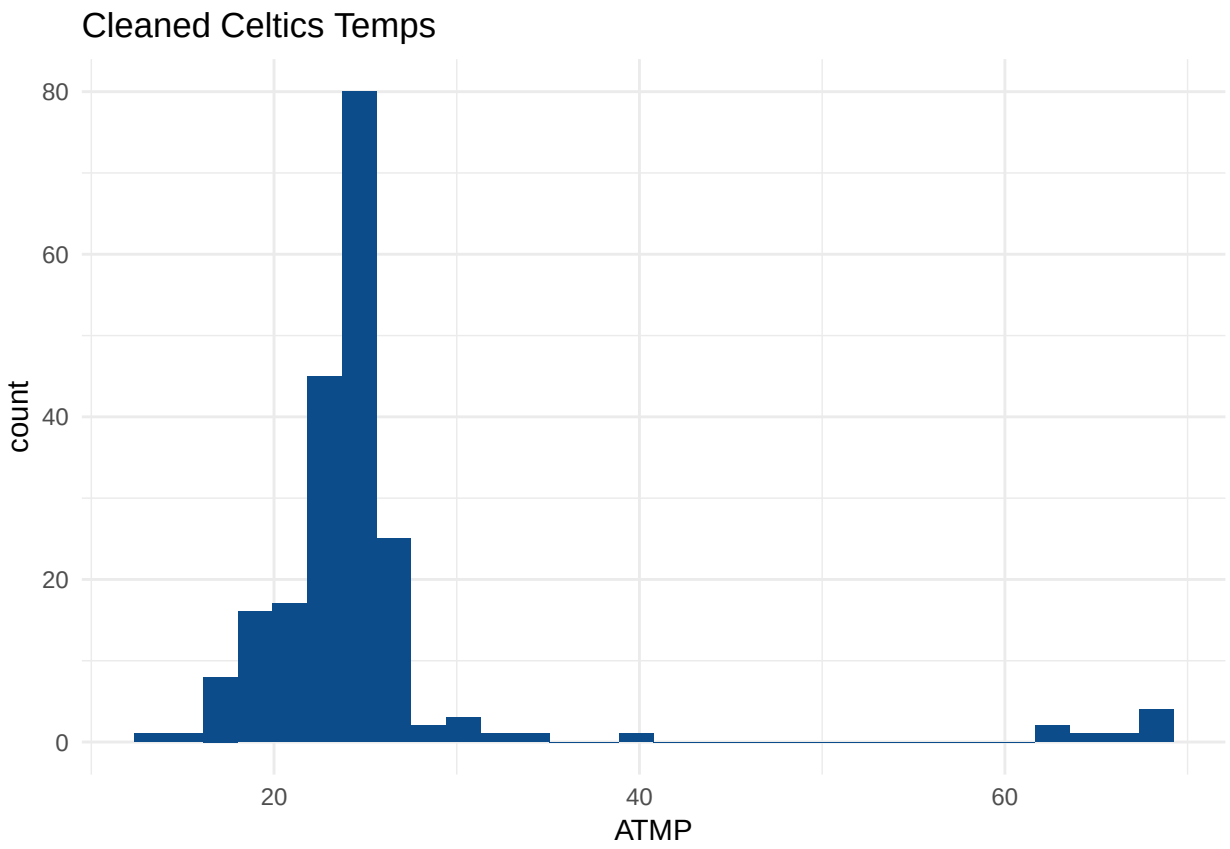
Overview

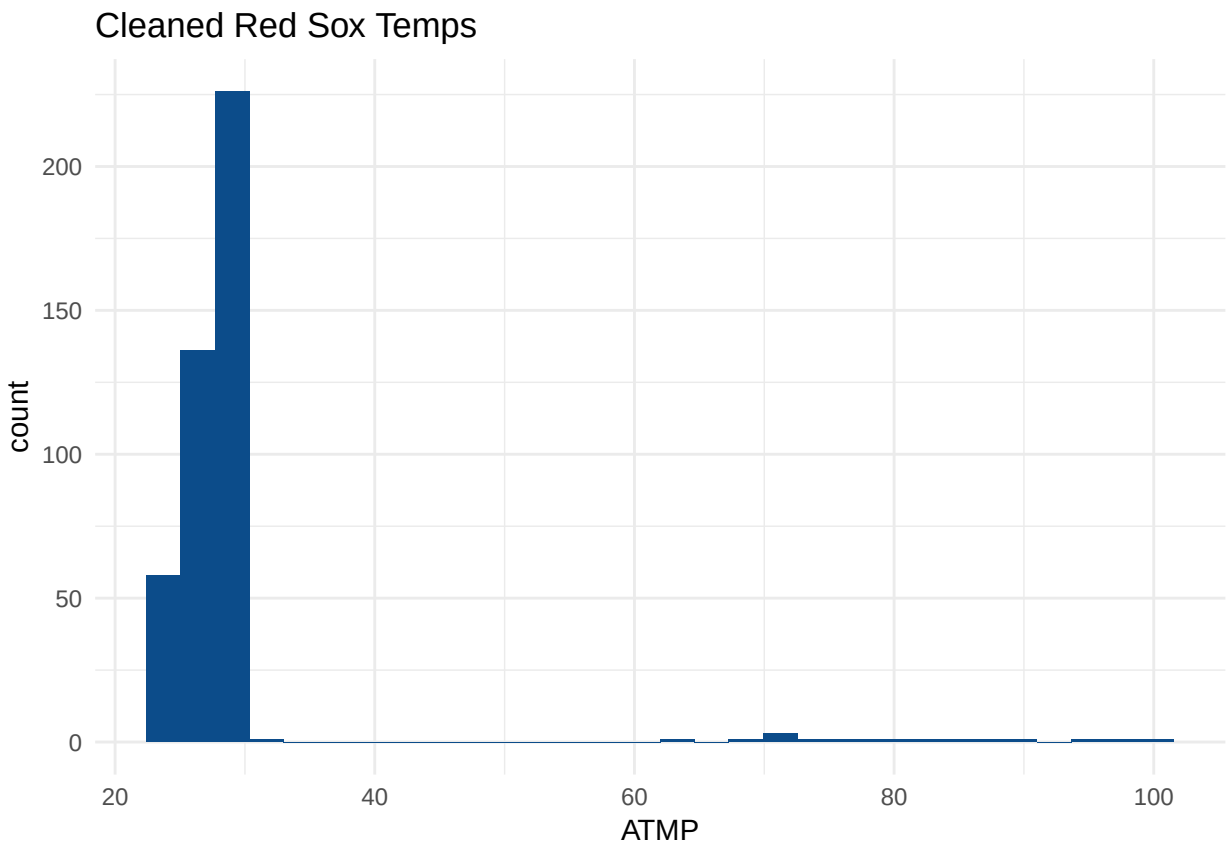
To make sure there is no abnormal data lies in the dataset, we first made graph to see the distribution of air temperature.


Celtics Temperatures
80
60
40
20
0
count
20
40
60
ATMP



As shown in the above graph, we found some abnormal records of air temperature. First, there are couple records indicate the air temperature is over $104°F$, we consider these as abnormal temperature since there is not such high temperature has ever happened in Boston (the record high temperature in Boston is $104°F$, recorded July 4, 1911). We also found lots of temperature records were at $70°F$ to $104°F$. Since most of the Celtics games are in Fall/Winter and early Spring, we thought these records could not represent the actual temperature in Boston area during those game days. We found the average temperature in Boston in October is $62°F/47°F$, and average temperature in April is $56°F/41°F$. Therefore, we set the threshold at $70°F$, hence we remove the records that has temperature higher than that.

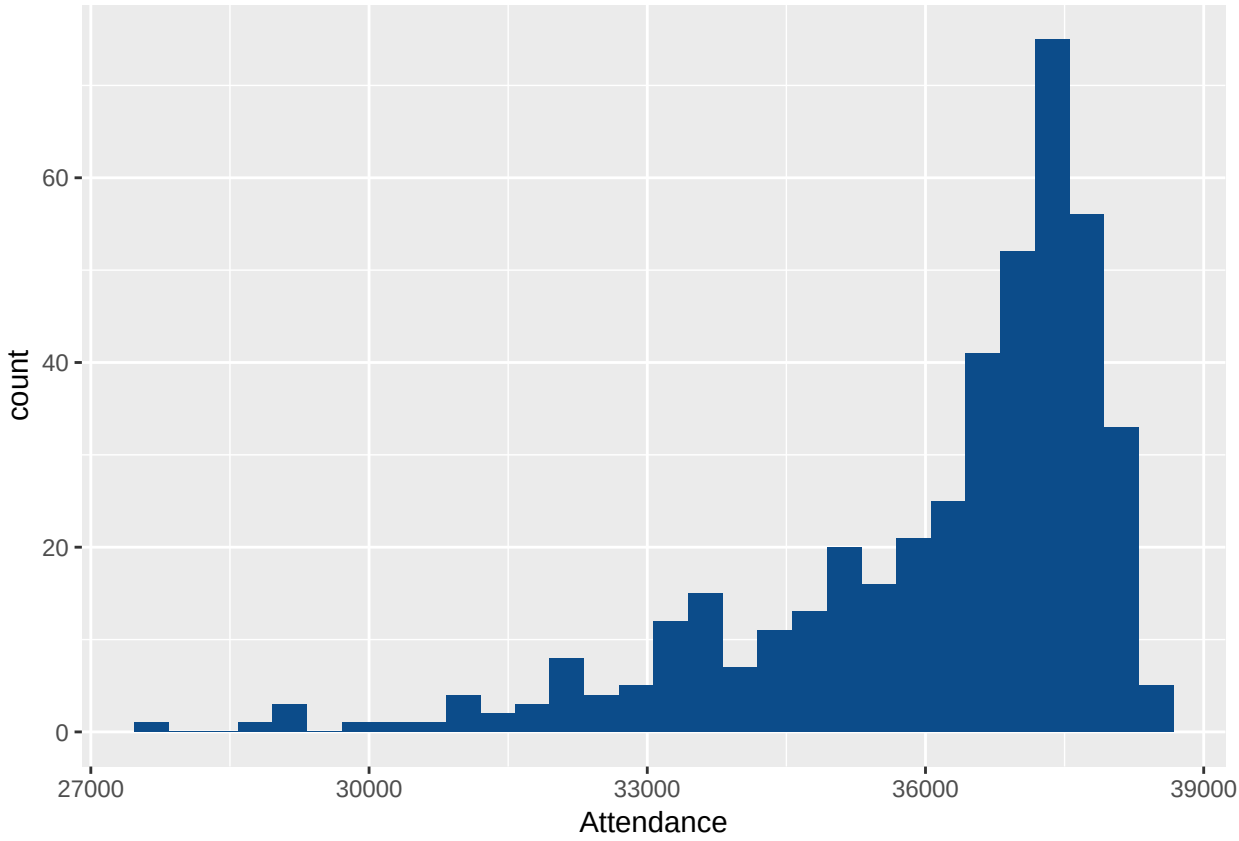

Cleaned Celtics Temps
80
60
40
20
0
count
20
40
60
ATMP



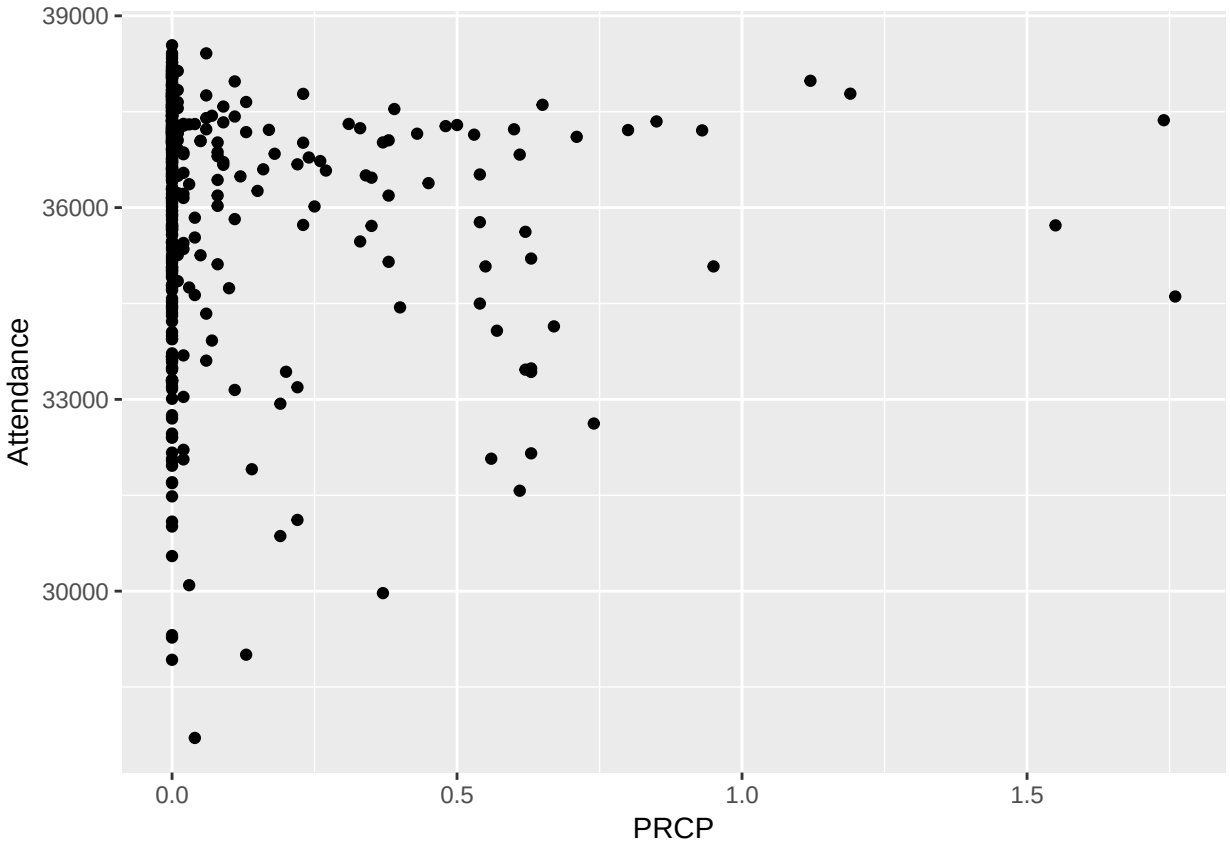
As we explored the temperature distribution of Red Sox games, we found extremely high temperature of $999°F$ on June 16, 2016, $675.7°F$ on July 25, 2016 and so on. Since the record high temperature on the earth is $134.1°F$, we seriously consider these records invalid. As we stated before, the record high temperature in Boston is $104°F$, we set our threshold here, and we removed the records higher than that.

Red Sox Attendance

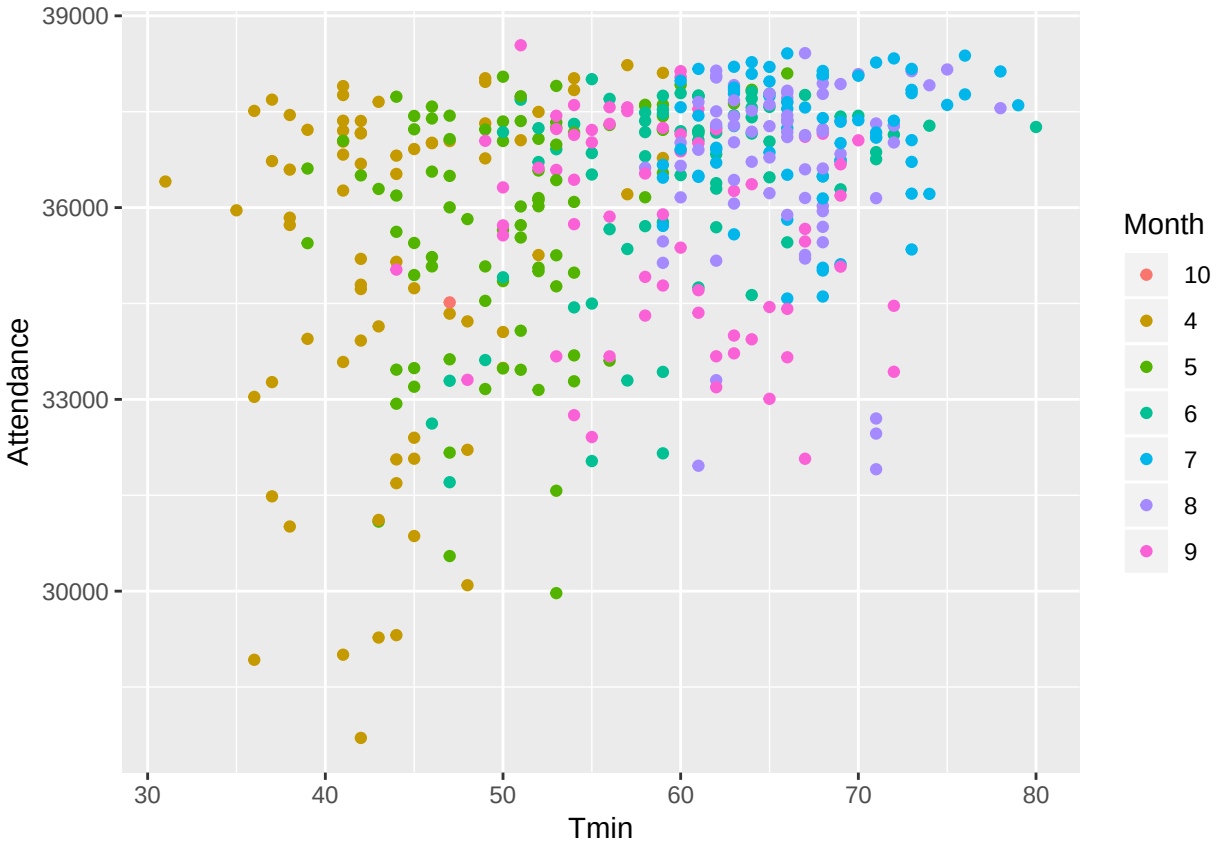
Creating a histogram of RedSox attendance, we can see that most games have pretty high attendance. Baseball is interesting because they sell "standing room only" tickets that can lead to attendance being higher than actual seats in the stadium. We can see this effect of especially popular games on the right tail of our histogram.

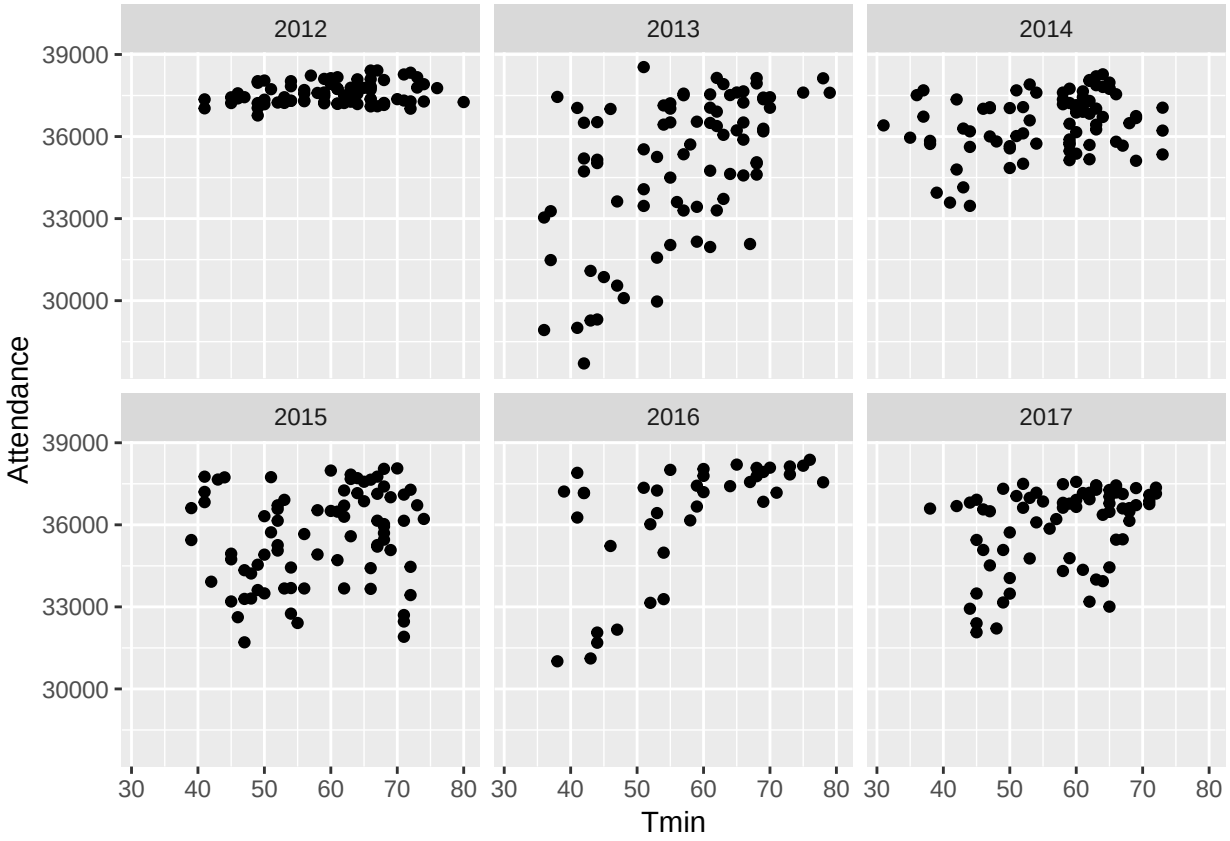


Plotting the rain vs. attendance, we see that there really isn't any significant influence. If there is a large rainfall, the games will be delayed or canceled. I am guessing many days with lage rainfall data had rain at a different time of the day from the game, which did not effect attendance.

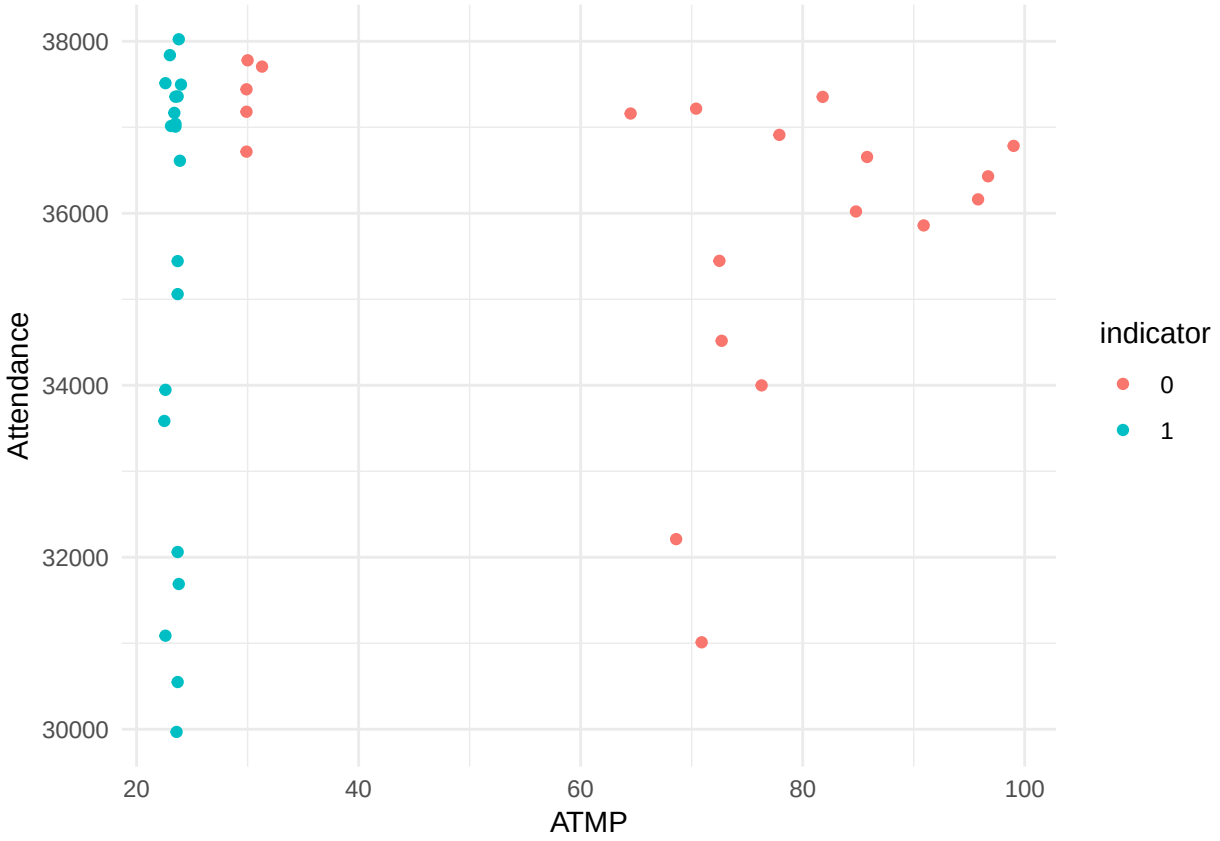


Temperature does seem to have an overall correlation with attendance. The first plot to follow shows minimum daily temperature vs. attendance, colored by month. Obviously the summer months (June, July, August) have generally higher temperatures and higher attendance. The second plot shows the effect that temperature has on attendance for each year. It is interesting that the different effects observed are not necessarily changed by how good the Red Sox are that year, but by how good they were the previous year. For example the Red Sox had a poor season in 2012, but there was a lot of excitement early in the year from a good 2011 season. The level of excitement around the team in April determines this variability since April is when many of the low temperature games take place.

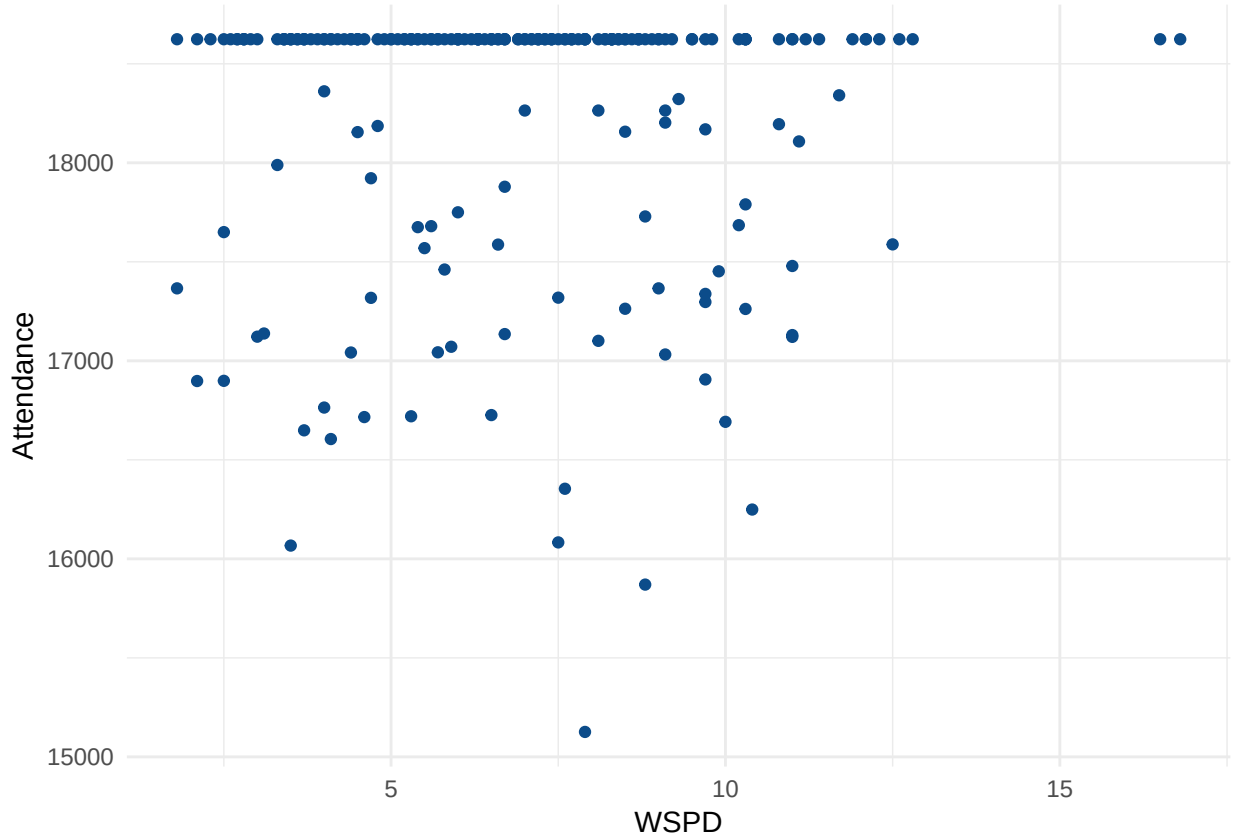

39000
36000
33000
30000
Attendance
30
40
50
60
70
80
Tmin
Month
10
4
5
6
7
8
9



We decided to plot the attendance for warmest days and coldest days to investigate further. Again, the relationship between attendance and temperature is still not significant. But we could tell some potential trends from the this chart. Below $30°F$, the attendance has a pretty obvious decrease. However, for the warm day games, there is no obvious relation between attendance and temperature.



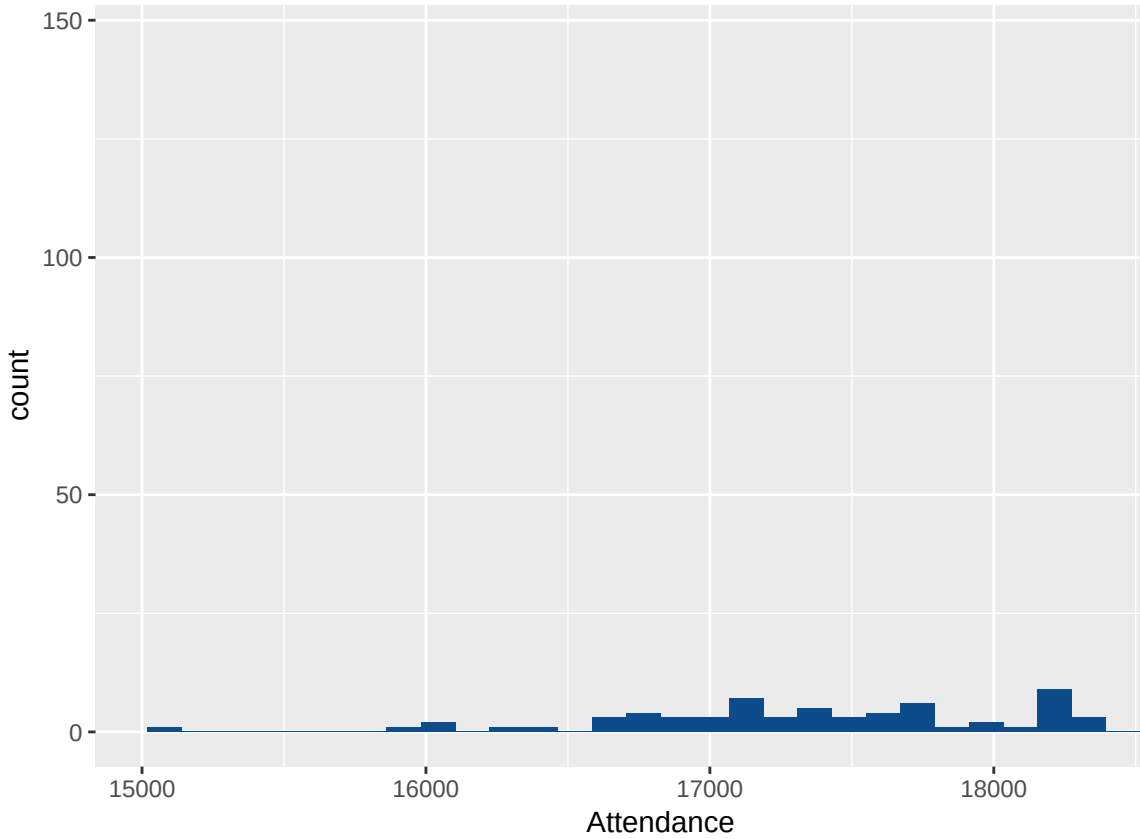
We also consider the level of wind speed could affect the attendance of Red Sox games, since baseball is outdoor and if there is strong wind, players??? performance may be affected and the quality of the game may be different as well. Thus, we assume less people will go to the game during windy days.



From the above graph, we can not tell the relation between attendance and wind speed. Even when the wind is really strong, the attendance is still very high. Since basketball is an indoor sports, therefore we didn???t assume wind speed will affect attendance.

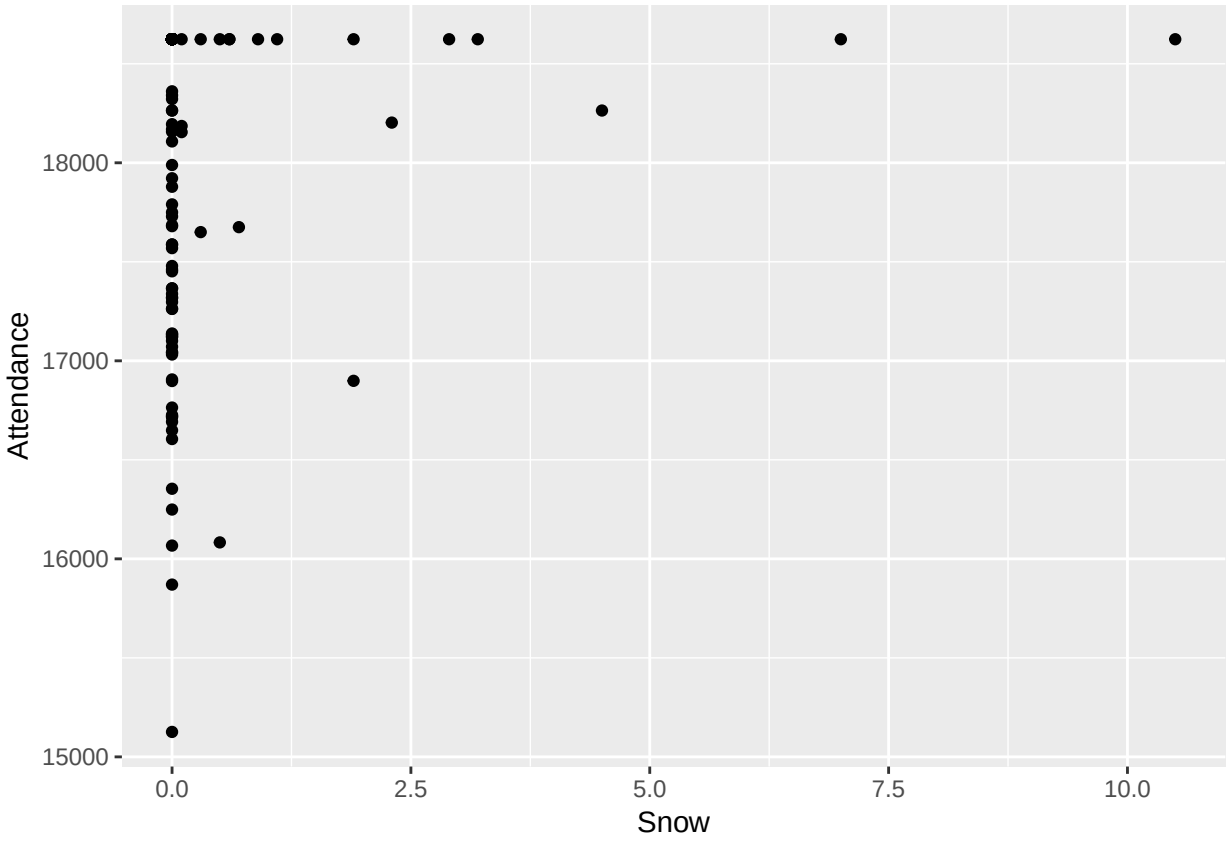
Celtics Attendance

As we did for the Red Sox, we can create a histogram of the celtics attendance. Unlike the Red Sox, the Celtics do not sell standing room tickets, so we see that they sell out many games which leads to the same attendance

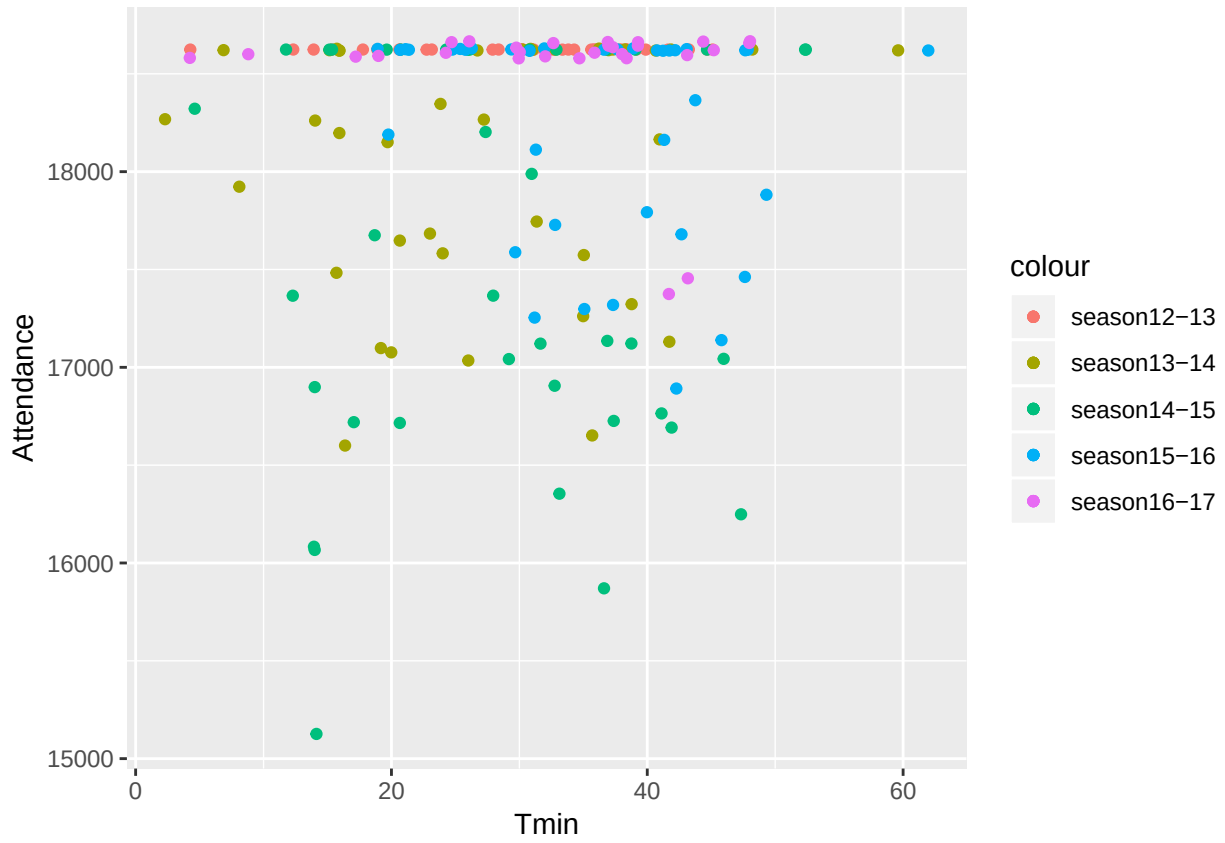


numbers for those games.

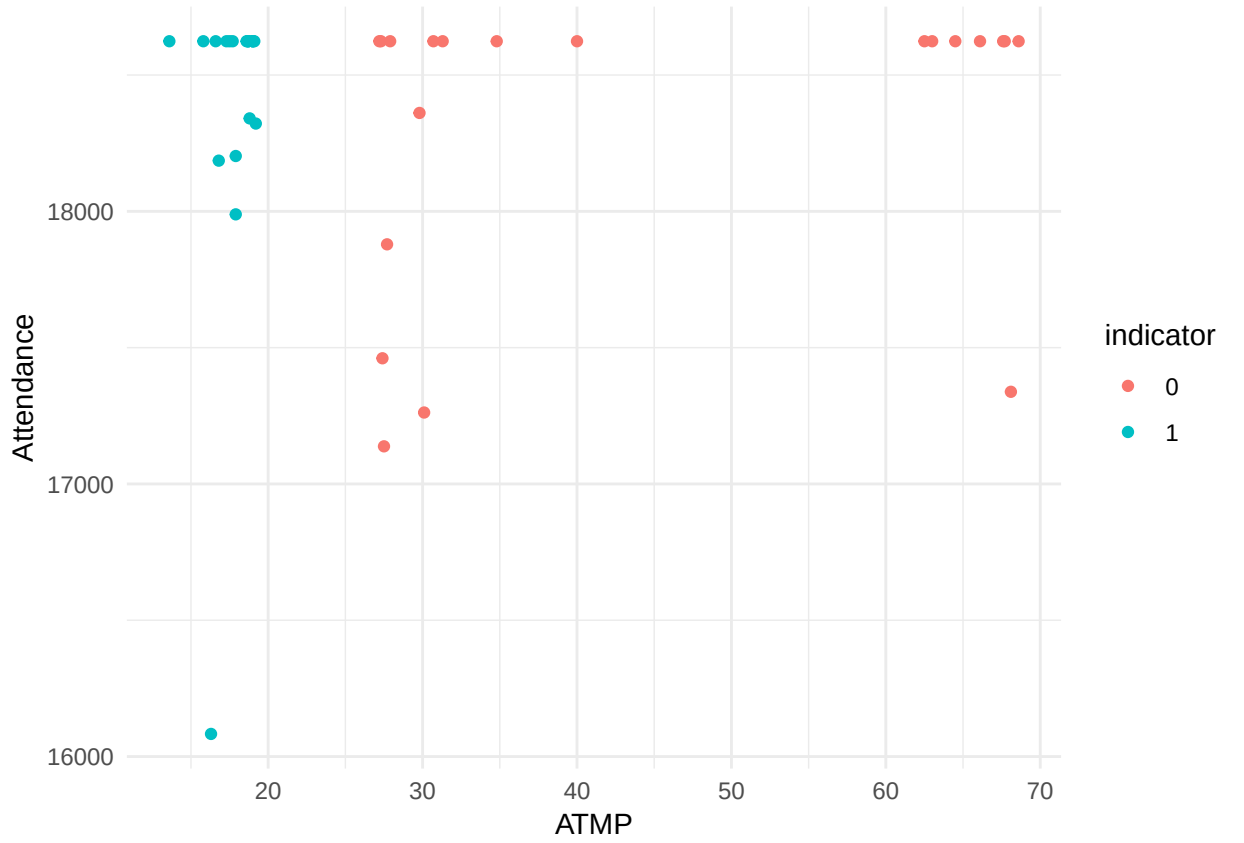
Since the Celtics play indoors, rain shouldn't effect attendance. Instead, we plotted snow vs. attendance to see if large snow storms would keep people from the games. The largest two storms didn't stop those crazy Celtics fans from supporting their team!



Temperatures also seem to have little to no effect on attendance, but by plotting the color of each point according to season, we can see that the success of the team does. The low attendance points are almost all from seasons where the Celtics struggled.



Again, we looked at the coldest and warmest days to further investigate effect of temperatures. Again, no real significance.



Interactive Data Visualization

In order to improve the user experience for people who look at our analysis, we created a ShinyApp for them to easily manipulate and explore the whole analysis in their preferred ways. In the ShinyApp, we include four drop down menus that allow users to choose specific date, variable, group and plot type. The scatter plot in the picture is an example showing the attendance by temperature. We hope everyone would stop by and enjoy our hard work.

Conclusion

Weather does not appear to keep people from attending sporting events. Warmer temperatures does lead to more people attending outdoor events, but team success and excitement about the team are most likely bigger contributing factors. Rain probably has some effect on people attending Red Sox games, but if it rains too much, the game is canceled. Also, we would need to investigate further to determine how rain impacts attendance depending on the time the rain occurs relative to gametime. The Celtics attendance, indoor sport, was not impacted by weather, even with large snow storms.