***The Impact of Star Power and Team Quality on NBA Attendance***

Henry Nketeh

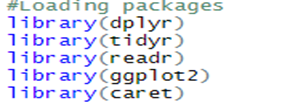
7/22/2019

**Introduction**

Data science and analytics has become more prevalent and has increased rapidly in the past decade. With this increase sports analytics is playing a more crucial role in the sports industry. Professional leagues such as the National Basketball Association (NBA) have utilized on court analytics to increase player and team performance, while management utilizes off-field analytics to get a better understanding or insight of the business side of their franchise. As a direct indicator of ticket sales, attendance is an important area to study. Many factors affect attendance, but the influence of each factor is different. The purpose of this research project is to determine the team rating a team require to guarantee a 1% home game attendance increase of 179 people per season.

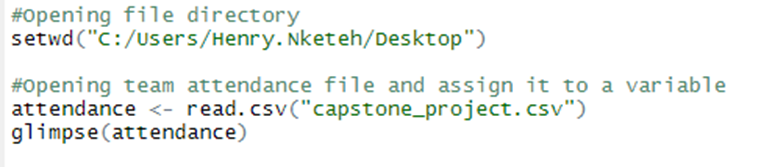
Information from my project will be helpful for the NBA owners and other professional and semiprofessional leagues. The clients that this project will deal with will begin with NBA sport teams then towards other leagues such as football and etc. Clients will be interested in this project because of how the potential increase in value can attract larger big ticket players to their organization. The data information involved in the project includes the 2018-2019 NBA teams home game attendance data, the 1% increase of the home game attendance, the percent of attendance increase, the team rating, and the net rating (NRTG). The approach that will be taken in regards of solving the problem is after analyzing all the data gathered and then data wrangling it. We then created the best linear regression model to predict the most effective team rating required for teams to pass that 1% attendance increase threshold. The deliverables of the project will be presented on a blog summarizing everything that the project contains.

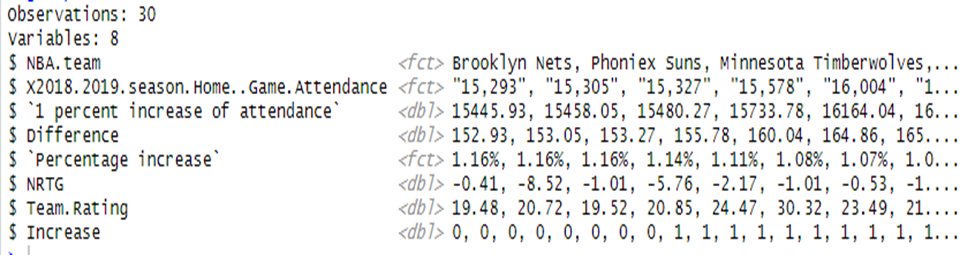
Let us start off by declaring the libraries we would be using in our analysis. The packages dplyr, tidyr, and etc. are great package to have as it contains virtually every tool that we would require for data wrangling.



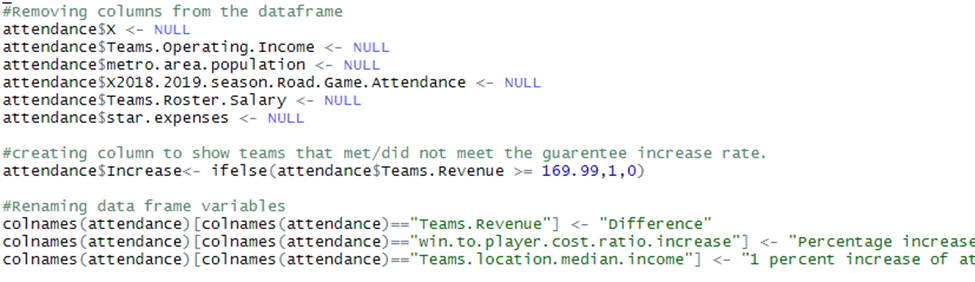
**Exploratory Data Analysis**

We started off by importing our data so that we can utilize it.

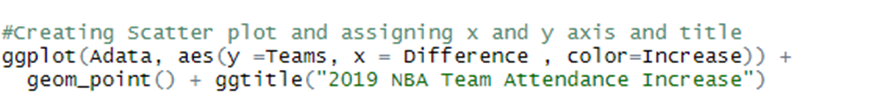


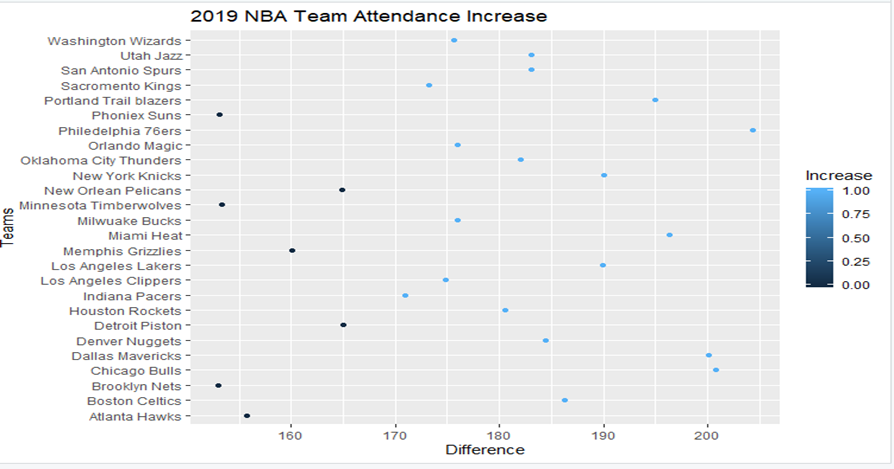


As you see there are 30 observations (rows) and 8 in our dataset. All of the variables are very self-explanatory but there are a couple variables that need to be renamed, deleted, and added another variable to display 1 or 0 for the teams that did or did not pass the 179 average people threshold after the 1% attendance increase. We can easily manipulate the data to fit our needs by utilizing the dplyr package.



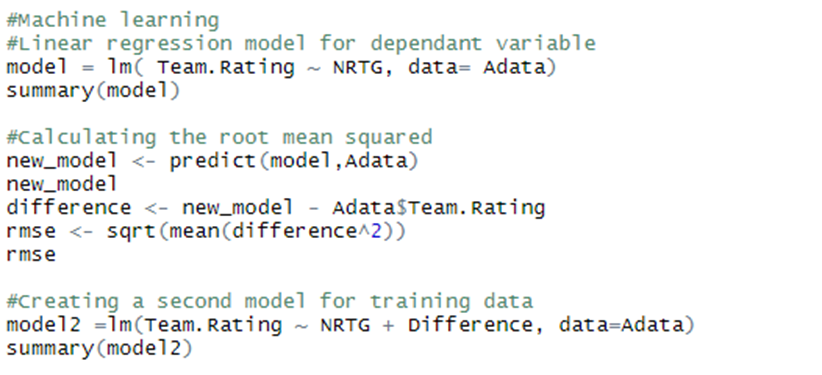
We can also view the data as a scatter plot.



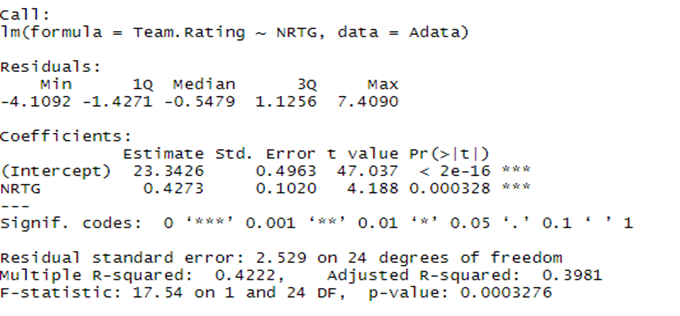


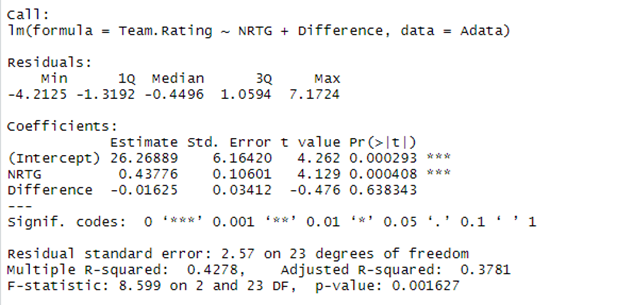
The x-axis contained the NBA teams while the y-axis contained the 1% increase home game attendance and home game attendance differences per team which was the Difference variable.

With this graph you can see that the Phoenix Suns, New Orleans Pelicans, Minnesota Timberwolves, Memphis Grizzlies, Detroit Pistons, Brooklyn Nets, and Atlanta Hawks did not pass the 1% threshold. Furthermore into the project we soon began to realize a couple of limitations that our project possessed. The limitations our project possessed is that there was a lack of independent variables that could support the dependent variable. As we continued gathering our data statistics we began constructing our linear regression model to determine the specific type of variable model to use, either it be a one or a multiple variable model.



The model measured the p-value, multiple R squared value to determine the significance of our independent variables on the dependent variable team rating.





After testing and examining the significance levels of the independent variables, we found that the model consisting of just the NRTG variable showed that it presents a significant impact on the team rating with a p value of .0004916, a three star significance code equating to 0, and R squared value of .4033 meaning that it is very significant. But when creating another model consisting of both independent variables and viewing the results it proved that the Difference variable significance code fell below three stars with a code equaling up to 1 with no stars, a p value of .002637 showing that there is little to no significance of that variable. Furthermore after examining the results we decided to approach the problem with a one variable model due to only one of our variables being significant.