Alex Yom

Capstone Project #2

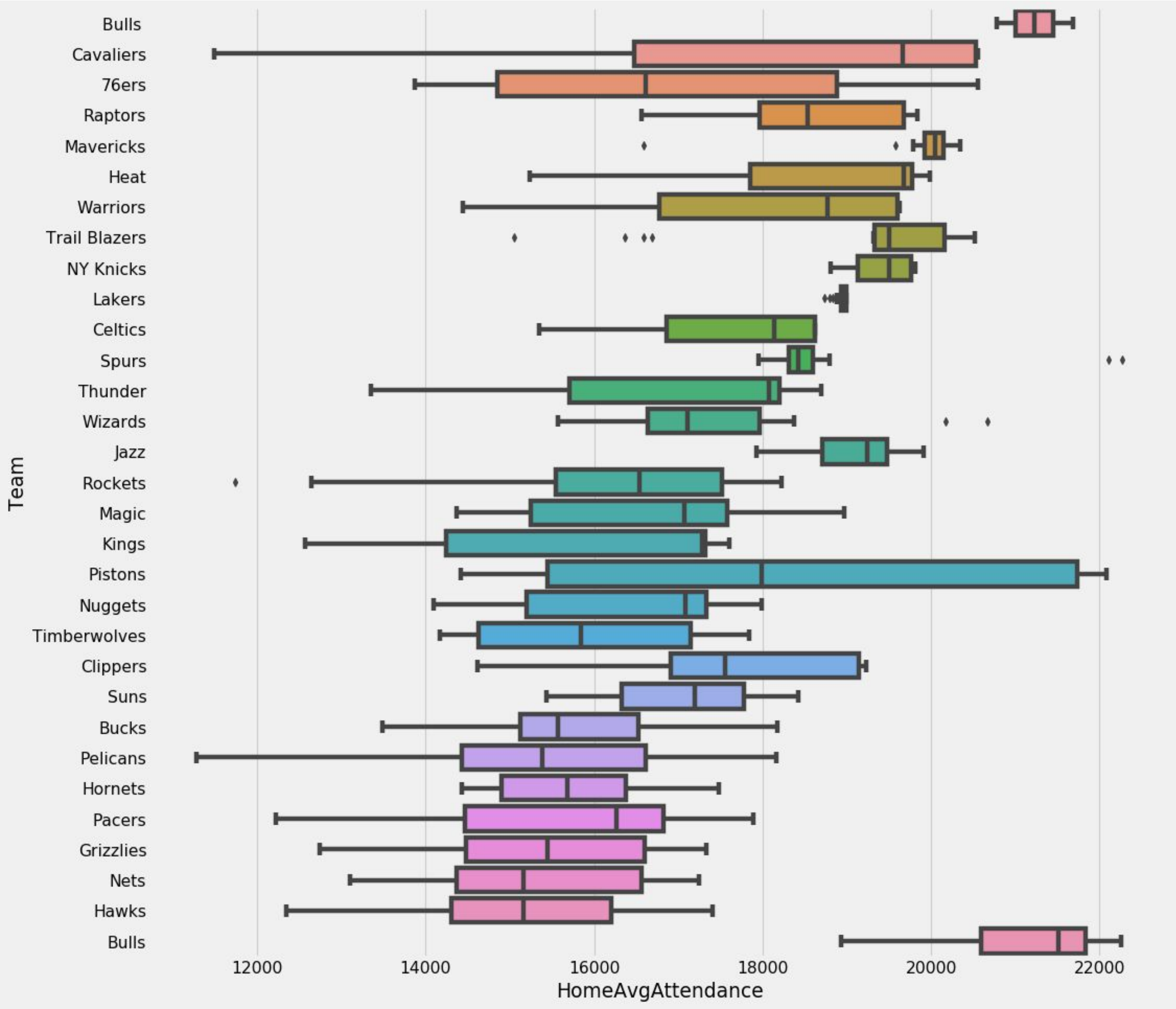
In sports, having the home-field advantage can be the difference between a win or a loss. But at the same time, having a successful winning team may be a prerequisite to have a cheering, large home crowd. Team owners would also love to sell out every home game to maximize revenue and boost team morale. What factors separate the most popular teams from the lower echelon teams? Are people more inclined to fill the seats of the stadium if their home team wins more often? If it can be proved that wins on the court can lead to fans in the seats, it would be in the best interest of team owners and basketball executives to assemble the best teams they can.

This project will explore a dataset that chronicles the ranking of every NBA team based on average home attendance starting from the 2000-2001 season to the 2015-2016 season. The team are ranked from highest average home attendance to the lowest. The teams are ranked from 1-30 with 1 being the team with the highest average attendance. The dataset also includes the number of wins that each team accumulated throughout the season to see if winning had a correlation to higher attendance. For a more in-depth analysis, I plan to add more features to this dataset by combining it with another dataset that will be assembled. Instead of just comparing the raw average numbers of attendance, I plan to compare the percentage of the arena that each team sold per year and compare the percentages. Since each team plays in a different home arena which have varying capacities, it would be more uniform to compare the percentages. By computing the percentage of arena filled and inputting those values to a new column, we would be able to compare the teams with a more unbiased perspective. Teams that play in smaller arenas will not be put in a disadvantage and teams with larger arenas will not have an advantage. Another column depicting whether or not the team made the playoffs that year will be added to append another measure.

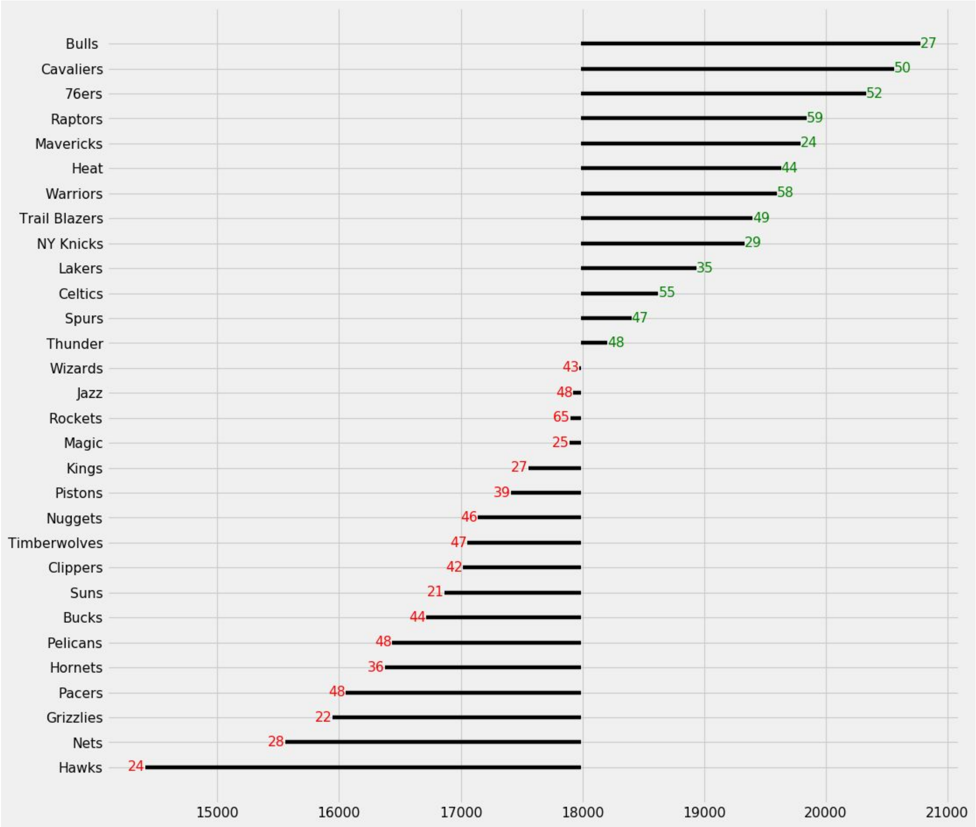
I plan to use data visualization techniques to illustrate correlations and trends to represent the data on individual teams and how they rank in the league and year by year. I plan to also use machine learning models to possibly predict or speculate how each team would have to perform in order to maximize the percentage of the arena that is filled. It will be interesting to see the different trends for each team and how they rank in terms of percentages of filling their home arenas.

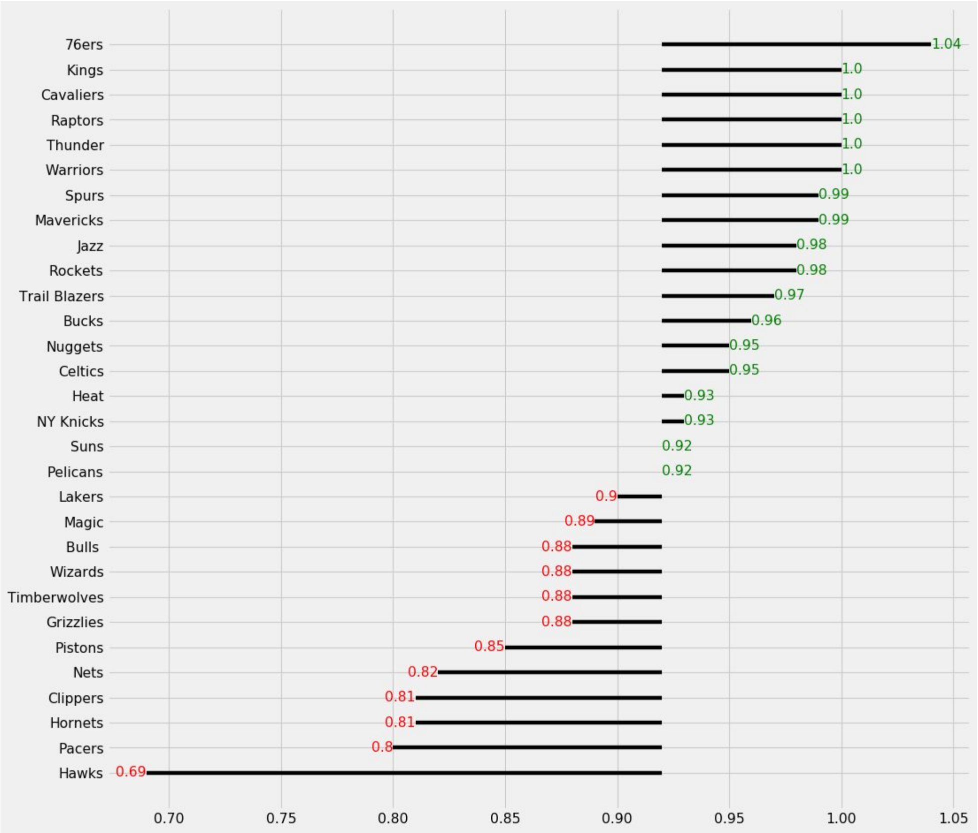
The data wrangling and cleaning process consisted of many steps due to the fact that three different datasets were being concatenated and merged together. The biggest data consisting of the data from years 2000-2015 served as the foundation because it was the largest dataset and had already been cleaned. The new dataset had the information from the years 2016-2017 and had to be merged with the foundational dataset. The values of the dataset were converted from integers to floats in the transition from excel to the notebook so the values had to be converted back. To ensure that these two datasets could be merged without complications, some columns were renamed and other columns were dropped. Lastly, the dataset containing the maximum capacity of every individual team's arena had to be merged with the dataset. After merging all three datasets, it was time to inspect the the dataset and make sure all the values were present and all values were the correct data type. During the merge, some values did not concatenate correctly and resulted in Nan values so those values had to be located and restored to their correct values. The final step in making sure our dataset was ready to be explored was the addition of a brand new column. The 'HomePercentCapacity' column would be the new feature added to this dataset. This column would take the percentage of the 'HomeAvgAttendance' of each team and divide that value over the 'HomeCapacity' value of each team. The new column would give us the percent that each team filled up their home arenas for that season.

Some preliminary data exploration revealed interesting information.



The boxplot above shows the culmination of the average home attendance for each team throughout the 17 seasons. There are several findings that this boxplot can highlight, that would not have been easy to see just by looking at the dataset. Some teams experience a great spectrum in terms of their average home attendance. Teams such as the Cavaliers, Pistons, and the 76ers all exhibit large disparities in their home attendance while other teams such as the Lakers and Spurs are amongst the league's most consistent in terms of average home attendance. There is little change in their attendance. There is a lot of information displayed on this one graph and each team can look at their own box plot and see the variance in their attendance. Some teams experience no variance in their attendance through the years while some teams experience huge waves of variance in their attendance. The one other feature that fluctuates or remains consistent year by year is the amount of wins for each team.


Bulls 27
Cavaliers 50
76ers 52
Raptors 59
Mavericks 24
Heat 44
Warriors 58
Trail Blazers 49
NY Knicks 29
Lakers 35
Celtics 55
Spurs 47
Thunder 48
Wizards 43
Jazz 48
Rockets 65
Magic 25
Kings 27
Pistons 39
Nuggets 46
Timberwolves 47
Clippers 42
Suns 21
Bucks 44
Pelicans 48
Hornets 36
Pacers 48
Grizzlies 22
Nets 28
Hawks 24
15000
16000
17000
18000
19000
20000
21000


76ers 1.04
Kings 1.0
Cavaliers 1.0
Raptors 1.0
Thunder 1.0
Warriors 1.0
Spurs 0.99
Mavericks 0.99
Jazz 0.98
Rockets 0.98
Trail Blazers 0.97
Bucks 0.96
Nuggets 0.95
Celtics 0.95
Heat 0.93
NY Knicks 0.93
Suns 0.92
Pelicans 0.92
Lakers 0.9
Magic 0.89
Bulls 0.88
Wizards 0.88
Timberwolves 0.88
Grizzlies 0.88
Pistons 0.85
Nets 0.82
Clippers 0.81
Hornets 0.81
Pacers 0.8
Hawks 0.69
0.70
0.75
0.80
0.85
0.90
0.95
1.00
1.05

The two graphs above, paint an interesting distinction between wins and filling up the seats. The first graph shows each team in terms of average home attendance in relation to the league average attendance. The numbers represent the number of wins that each team had during the season and the red coloring means that the team is below the league average while the green means that the team is above league average. Some interesting points to note are that the bulls, despite winning a measly 27 games in the 2017-2018 season, led the league in average home attendance. The mavericks and knicks also placed in the upper echelon of attendance despite winning less than 30 games. On the other hand, the rockets actually led the league with 65 wins, were below average in terms of attendance. But, the second graph paints a different picture and puts some perspective on reality. The second graph shows the same data as the first graph except it shows the percentage of the maximum capacity that the teams averaged for the season. When compared side-by-side, some teams are repositioned. The kings who were below average in terms of average attendance strictly by the numbers, actually averaged a sell-out crowd every home game, meaning their stadium was completely full for their home games. The bulls, who led the league in average attendance, dropped below average in percentages as they came in at filling 88% of their seats on average.