What Are Some Predictors of Sporting Event Attendance?

By Jackson Downs

Introduction:

Sports are a way of life in the United States. The names LeBron James, Tom Brady and Mike Trout draw in hundreds of thousands of fans to their respective sporting events each year. Michigan regularly draws over 100,000 people to the big house on fall Saturday's and only 2 teams in the NHL (the Sabers and Coyotes) failed to fill their venues to at least 70% on average.

Attendance is also a huge generator of revenue. The Covid pandemic had a huge effect on team's revenues, as they became almost solely reliant on TV money. For college teams, money from college basketball and football teams fund other less profitable teams. While TV deals may be the main driver, attendance makes an impact as well.

And most importantly, a team's attendance gives them the home field advantage. Think of the Seattle Seahawks' 12th man. Notoriously difficult places to play such as Duke's Cameron Indoor Stadium or Boston's Fenway Park give teams advantages that are difficult to quantify with statistics.

Because attendance is so important, I wondered what factors influenced attendance. Did a team playing well lead to higher attendance? What about a more attractive playing style with more points being scored? Did older, historic stadiums entrance fans?

In this paper, I'm going to see if some of the questions are correct and if any of these variables influence attendance of six sports (NBA, MLB, NHL, NFL, D1 Men's College Basketball and FBS College Football).

Background and Literature Review:

Articles have been written on everything from fan satisfaction at PGA tour events to air pollution's effect on attendance in the Chinese Super League[1]. For years, people have been trying to decipher what factors bring people into different sporting events. Is there a universal factor across all sports? Or are there individual factors in some sports that play a role?

In my research, I found many articles across a wide variety of sports that attributed a lot of different factors toward attendance. One in 2009 talked about how emotional factors and facilities played a huge role in attendance[2]. Does this mean new stadiums and arenas will draw in higher attendance? There is the possibility I overestimated the prestige factor.

---

[1] Air Pollution and Attendance in the Chinese Super League: Environmental Economics and the Demand for Sport (researchgate.net)

[2] j.ijhm.2009.10.01120170108-1649-1sl7v1b-libre.pdf (d1wqtxts1xzle7.cloudfront.net):

Another article stated that loyalty and psychological involvement have a huge impact on attendance[3]. Loyalty is a difficult thing to quantify, but teams with more diehard fans such as the Oakland Raiders and Toronto Raptors tend to be difficult road environments even when their teams aren't as successful.

A third article I looked at said identity salience was an important factor in attendance[4]. In simpler terms, it means that fans flock to see the most well-known players. While I didn't include specific players in my analysis, it would be interesting to see the effect someone like LeBron James has on attendance at home and on the road.

Across all the articles I read, there were a multitude of other variables tested such as weather or game start time[5]. However, articles directly looking between winning, and attendance weren't as common. I believe this was because there is a general belief that teams that win draw in fans. Outside of super popular teams like the Red Sox or Cubs, teams that lose won't draw in as many fans. We'll see if this is true in our analysis.

My data and variables (and why I chosen 4 main variables and dependent variable):

In my analysis, I used a main dependent variable and four predictor variables. Additionally, the MLB I used a fifth which I'll get more into.

Additionally, the Excel data for all of this is linked here: Attendance Analysis.xlsx

Dependent Variable:

Percentage Capacity: Not all stadiums are created equal. A basic attendance number may be unfair because stadiums vary in size. Fenway Park is way smaller than Dodger Stadium so the capacity at which a stadium is filled is a better representation. This is especially true for college as many small Division 1 schools have very small arenas. This is the reason I chose Percentage Capacity as my dependent variable.

Before we move on, one thing to note. Some college basketball teams played in multiple arenas such as Villanova and St John's, so their attendance could be skewed based on playing in a bigger arena. However, I decided that with over 350 college basketball teams, the impact wouldn't be impactful.

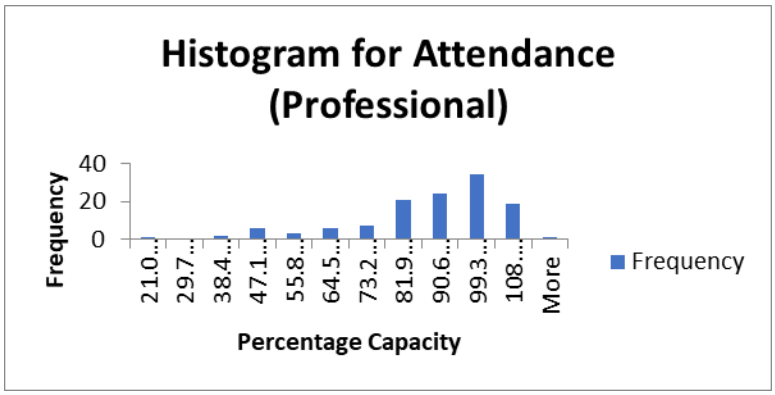
In looking at Percentage Capacity, the mean for the 4 professional sports and 2 college sports was 57%. This felt low, so I compared it to the mean for only 4 professional sports which was 83.6%. Additionally, the only professional sports median was 88.6% so some really bad teams were weighing it down. Overall, most professional teams have very high attendance capacity while it varies way more in college. You can see the difference in the two histograms below.

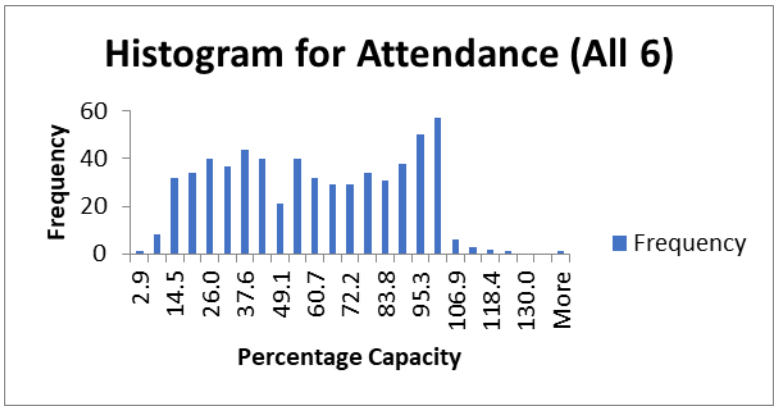
---

[3] Repeat_Attendance_as_a_Function_of_Invol20160402-19341-1wazkox-libre.pdf (d1wqtxts1xzle7.cloudfront.net)

[4] jlr-volume-32-number-2-pp-225-246.pdf (nrpa.org)

[5] indice (researchgate.net)



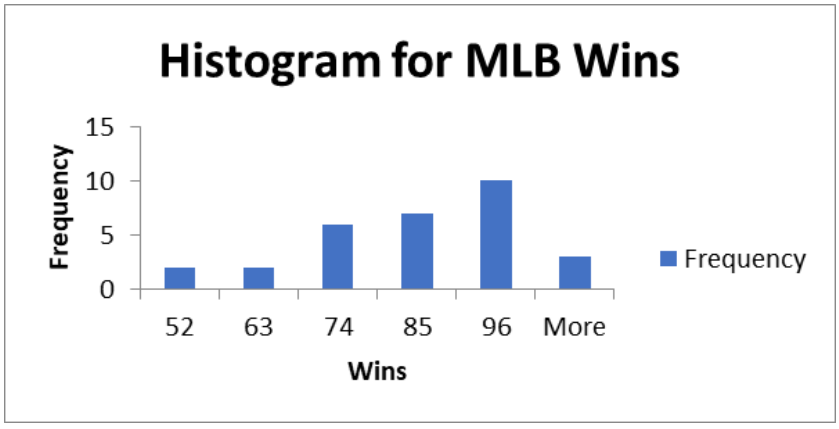


As you can see college is a lot more uniform, there are teams that fill up their arenas and stadiums consistently, others that do partly and some that struggle to get anyone. In professional sports, 80% is almost the bare minimum most of the time.

Predictor Variables:

Team Wins: This statistic is the easiest to explain as it's the number of games a team won over the course of their season. For professional sports, this doesn't include playoffs, but it does for college sports as that's how their win loss records tend to be calculated.

Across all sports, wins tend to carry a bell curve like structure due to their nature. When two teams play, only one can get a win. Below, you can see a histogram charted for MLB Wins:

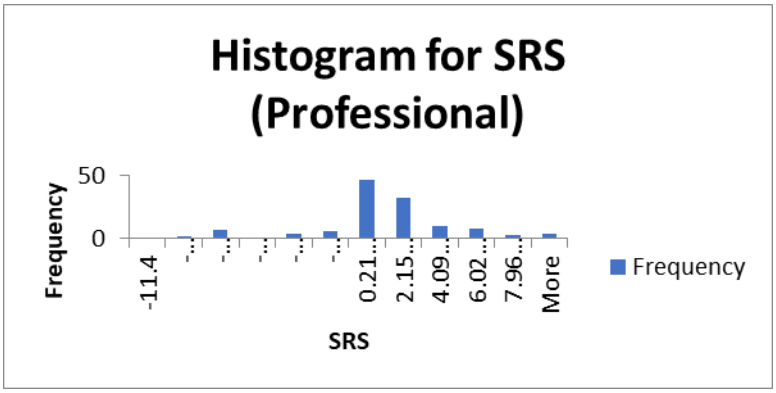


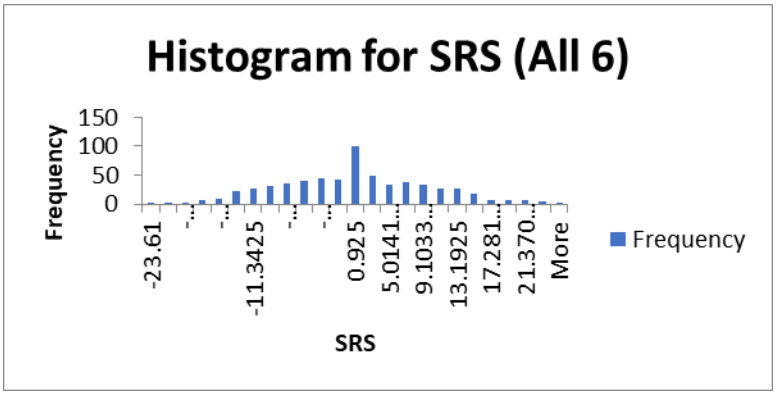
The histogram is slightly skewed left, and I found this to be a trend as the NBA had the same trend. I believed this to be because the bad teams bottom out, going all in on a rebuild while some teams toward the middle push toward a playoff spot. Even with this slightly skewness, you can see the bell curve shape for the most part.

SRS: Simple Rating System (SRS) is a statistic to compare teams based on their point differential and strength of schedule[6]. In the footnote, you'll be able to see how it's created more in depth. The important thing to know is the higher the number, the better the team. SRS exists on all the Sports Reference's and is supposed to be a universal way to compare teams.

In comparing only professional leagues to all six leagues, the descriptive statistics and histograms show a common trend. The mean hovers around 0 (with it being 0 in the professional leagues and -0.15 in all six). However, all six leagues have a much higher standard deviation of 8.9 to the professional league's SD of 3.9. This is unsurprising, given that there are three times the number of teams playing D1 College's Basketball than in all four professional sports leagues combined, but still important to point out.

Finally, we can look at the histograms and see a common bell curve between the two. The professional leagues isn't perfect, but you can see the majority of the teams have a slightly above average (above 0) SRS.





[6] SRS Calculation Details | Sports-Reference.com

Stadium Age: This statistic refers to the number of years since the stadium was built. My original thought was the older stadiums such as Lambeau Field would capture more attendance because of its historical draw. However, new stadiums like Sofi Stadium and Globe Life Park, should draw fans in because of its amenities.

I wasn't sure how the data for stadium age would look, but here are the summary statistics below:

Stadium Age - Professional	
Mean	23.90323
Standard Error	1.598436
Median	22
Mode	22
Standard Deviation	17.79943
Sample Variance	316.8198
Kurtosis	8.674631
Skewness	2.449247
Range	109
Minimum	0
Maximum	109
Sum	2964
Count	124

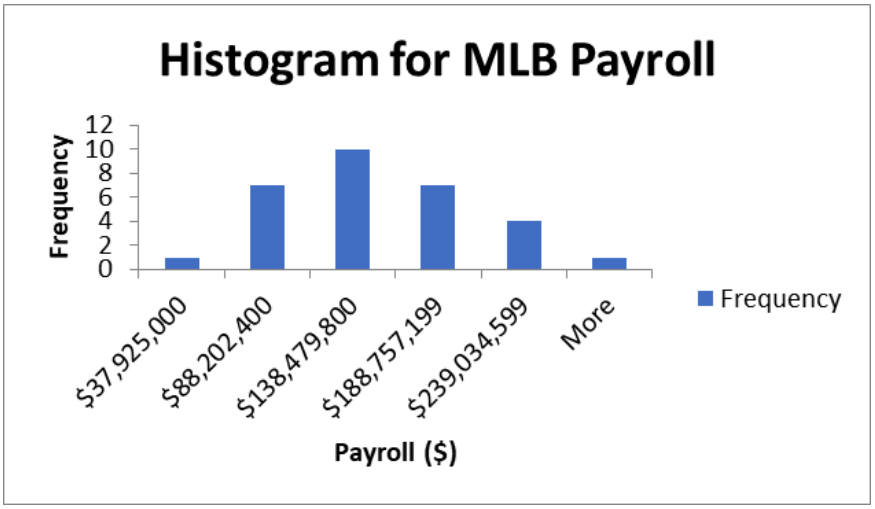
Stadium Age - All 6	
Mean	40.72951
Standard Error	1.039758
Median	38
Mode	22
Standard Deviation	25.68012
Sample Variance	659.4686
Kurtosis	-0.06522
Skewness	0.707737
Range	112
Minimum	-1
Maximum	111
Sum	24845
Count	610

As we can see, college arenas are 17 years older on average. A reason for this may be that a lot of college teams aren't trying to maximize attendance. They are happy with their on-campus gym and it would be a useless financial hassle to try and extend.

Only eight Professional stadiums are older than 50 years old with five being MLB stadiums, two being NFL stadiums and the other being Madison Square Garden. Nine College Basketball Arenas are older than 88 years old. In my analysis, it'll be important you remember the large disparity in stadium age if the variable is significant.

Points/Points/Goals For: This statistic shows how many points, goals or runs a team scored over the course of the season. The logic is that a team that scores more will be more exciting to watch, so they'll have higher attendance. I'm not going to show any descriptive statistics for this data as scoring varies between sports, but I would anticipate the data would look like a bell curve for each sport.

Payroll: Payroll is the estimated amount of money an MLB team is paying its players. In the NBA, NFL and NHL, there are strict salary caps that limit how much teams can spend to keep competitive balance. However, there is only the luxury tax to deter teams from spending too much, not stop them. Because of this, payroll tends to only matter in the MLB. This statistic was in the same dataset as attendance, so I decided to see if it can be a predictor for MLB attendance.



As you can see, Payroll mimics a bell curve as some teams pay a lot, some a little and most fall somewhere in the middle.

Adjusted R Squared

What is Adjusted R Squared?

Adjusted R Squared is a modified version of R-Squared that accounts for predictors that are not significant in a regression model[7]. R-Squared is used to explain the degree to which predictor variables explain the variation of output variables. In our analysis, R-Squared would be based on how Wins, SRS, Points/Goals/Runs For, Stadium Age and Payroll explain the variation in Attendance Percentage Capacity. If R-Squared is 0.9, it indicates 90% of the variation in Attendance Percentage capacity can be determined by the five predictor variables I just mentioned. The higher the R-Squared, the more your predictor variables explain the output variable. For Adjusted R-Squared, it shows whether adding additional predictors improves a regression model or not.

Why Do We Use Adjusted R Squared?

If you add more variables you add to a regression model, the R Squared will stay the same or improve because of how it's calculated mathematically[8]. This means that even if there's no relationship between the predictor and output variable, the R-Squared may improve.

This is where Adjusted R Squared comes in. Adjusted R Squared is adjusted for the number of predictors in a model. Therefore, adding more variables doesn't automatically make a model look better.

---

[7] Adjusted R-squared - Overview, How It Works, Example (corporatefinanceinstitute.com)
[8] How to Interpret Adjusted R-Squared (With Examples) - Statology

What is a good Adjusted R Squared Value?

A good Adjusted R Squared value is a difficult question because it depends on the field and opinion. In my analysis today, I'll be using tiers to separate adjusted R Squared. In these tiers, I'll consider any model with an Adjusted R Squared of below 0.3 insignificant. This is due to my past experience in analysis, but the number does tend to move around a lot based on the industry and data.

Analysis

Finally, we'll get to our analysis of the models. I have separated the models into three different tiers based on their Adjusted R-Squared. The first tier will be models over an Adjusted R-Squared of 0.5, second is 0.3-0.5 and lastly is models below 0.3. In each of these sections, I'll talk about the models and their relationship with Attendance Percentage Capacity.

Tier 1: Strongly Explains Attendance Percentage Capacity (0.5+ Adjusted R Squared)

Models:

- Only MLB Payroll (0.59)
- All 4 Variables - NCAAF (0.55)

Explanation:

Two of our models had an Adjusted R Squared of 0.5 or above, which is encouraging because that show they explain Attendance Percentage Capacity very well. The first model is MLB Payroll which had an Adjusted R Squared of 0.59. Basically, the higher payroll an MLB team has, the more likely they are to draw a high attendance percentage capacity. One thing to note is that payroll may correlate with the market the team is in. So, a team may have a higher attendance because they're in a big market and payroll is just a confounding variable.

The second model was for college football and included all four of our main variables. Since we used Adjusted R Squared, we know it wasn't buoyed due to having four variables. This leads me to believe that the combination of these variables explains Attendance Percentage Capacity well, especially with a sample size over 120. How each individual variable affects Percentage Capacity is a question to answer for later.

Tier 2: Mildly Explains Attendance Percentage Capacity (Adjusted R Squared between 0.3 and 0.5)

Models:

- CBB SRS (0.49)
- All 4 CBB Variables (0.48)
- NCAAF SRS (0.44)

Explanation:

Three models have an adjusted R Squared between 0.3 and 0.5, and they're all college sports models. Two of these models are college sports models involving SRS. The reason I believe this to be is that SRS takes into account margin of victory and strength of schedule. Strength of schedule has a far bigger disparity in college sports than professional sports due to conferences and the large number of teams. Therefore, bigger teams tend to have higher SRS's and therefore higher attendance. I'll touch on this in my conclusion, but most historically good teams tend to always be good because of their reputation. Because of this, they consistently draw in a good capacity of fans.

Tier 3: Does Not Explain Attendance Percentage Capacity (Adjusted R Squared Below 0.3)

Models:

- NFL SRS (0.29)
- NBA Wins (0.27)
- CBB Wins (0.26)
- All 4 NFL Variables (0.23)
- MLB Wins (0.23)
- NBA SRS (0.23)
- MLB SRS (0.2)
- NFL Wins (0.19)
- All 4 NBA Variables (0.18)
- CBB Points For (0.17)
- NCAAF Wins (0.16)
- NHL Wins (0.15)
- All 4 MLB Variables excluding Payroll (0.14)
- MLB Runs Scored (0.13)
- NHL SRS (0.13)
- NCAAF Points For (0.12)
- NCAAF Stadium Age (0.11)
- All 4 NHL variables (0.08)
- NHL Goals For (0.06)
- NFL Points For (0.05)
- NBA Points For (0.05)
- CBB Stadium Age (0)
- NBA Stadium Age (-0.02)
- MLB Stadium Age: (-0.03)
- NFL Stadium Age (-0.03)

Explanation:

A large majority of the models failed to even have an Adjusted R Squared of 0.3 showing they don't explain percentage capacity that well. NFL, NBA and MLB SRS's and Wins all had Adjusted R Squares above

0.2, showing that those statistics could be useful in some way to explaining Attendance Percentage Capacity. SRS may have flaws, but it was the best performing predictor variable out of the main four.

The NHL was by far the hardest sport to predict as no model had an Adjusted R Squared of above 0.15. A possible explanation is that hockey has some areas where fans show out more for various reasons. Nashville had an Attendance Percentage Capacity of 107.8% as they've become a Southern Hockey crazed city. Seattle was terrible in 2021 but managed a 100% capacity as the team is in its second season and Seattle embraces a new NHL team. Even teams that have huge fan bases that were pretty good the last few seasons had letdowns in 2021, I'm talking Las Vegas and Washington. Regardless, our models did not do a good job of predicting NHL Attendance Percentage Capacity.

The worst performing predictor variable seemed to be stadium age. In my analysis of Stadium Age earlier, I was mixed in whether the allure of older stadiums would draw new fans or the amenities of new ones. It turns out the explanation is a combination of the two. In all four major professional leagues and CBB, the Adjusted R Squared was between 0 and -0.03 meaning that it in no way explained Attendance Percentage Capacity. Even in NCAAF, the Adjusted R Squared of 0.11 isn't high enough to have any sort of weight Additionally, schools like Nebraska and Michigan could have skewed this. Overall, Stadium Age didn't explain Attendance Percentage Capacity in any way.

Conclusion and What Can be Improved:

In our analysis, we looked at four predictor variables of attendance percentage capacity: Wins, Points/Goals/Runs For, Stadium Age and SRS. We looked at these over six leagues, four professional and two colleges. From our analysis, I can conclude that these four variables only really predict attendance in the college leagues. Additionally, SRS seems to be the best statistic for predicting these college leagues.

In the future, there are many other variables that could be explored. I could use Win Percentage instead of Wins as that may be more accurate when comparing across sports. Other factors mentioned in academic journals like time of day and weather could play a role as well.

I'd also be interested to see how stadiums with good public transportation options influence attendance. Will people go if the stadium is easy and cheap to get too? Additionally, I'd like to explore at some point if star players make an impact. Do more fans come to the stadium when an MVP, #1 recruit or historic player comes to the arena? These are just some of the things I'd be interested in exploring.

Lastly, there's room for additional analysis factors beyond Adjusted R Squared. This was just a statistic that I was familiar with, but there are plenty of different data mining and data science techniques to predict attendance or analyze factors of attendance.