Syracuse University

Analyzing the Factors that Impact Attendance for the Brooklyn Cyclones

Alexander Borelli

SAL 213 M002

Dr. Losak

9 December 2020

Introduction:

For minor league baseball clubs, attendance is the driving force for bringing in a substantial source of revenue for themselves. Because of this, owners are always looking for opportunities and ways in which they can maximize their attendance from game to game. Ways in which organizations go about trying to increase attendance range from various promotions to improved team winning percentage. This paper will examine the numerous factors that drive attendance and will help decide which factors are most significant for the Brooklyn Cyclones attendance.

Summary of Data and Variables Used:

For our data collection, we used the MiLB website to retrieve game by game data from the 2019 Brooklyn Cyclones season. By doing this, I was able to create many variables for my attendance model. These variables include promotion dummy variables, day of the week, the month of the year, home winning percentage, weather, time of the game, opponent, and playoff games. We retrieved our weather data from Weather Underground which allowed us to potentially include temperature, precipitation, and humidity variables in our model. The Cyclones promotion variables were split into five categories: fireworks, discount, giveaway, bobblehead, and themed night. To be more specific on the discount and giveaway promotion categories, discount promotions were any night in which there was a sale on tickets or food and beverages and giveaway promotions were a night in which a product was given away to fans attending the game.

Attendance Summary Statistics	
Min	1428
Median	4452
Mean	4528
Max	8031
STD	1608.87

Temperature Summary Statistics	
Min	62.8
Median	75.9
Mean	78.35
Max	82.3
STD	4.66

Some data I wish I had but was not made available was whether or not a Major League player was playing a rehab game on a given day. A star player playing for a minor league team for a game is likely to increase attendance. For example, Aaron Judge playing a rehab game for Staten Island is very likely to increase attendance for that game.

Model Overview:

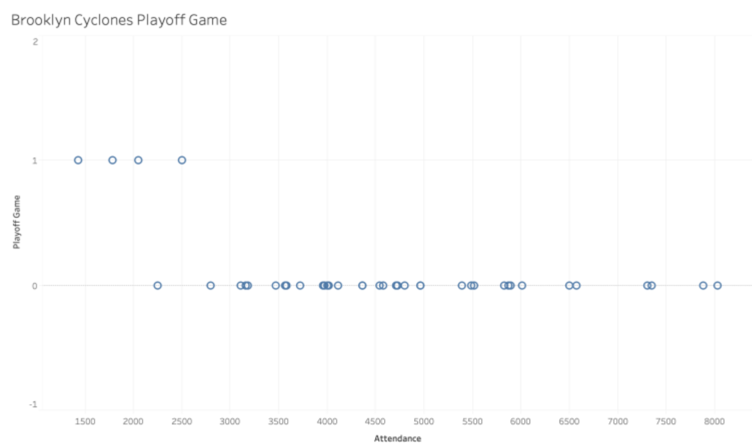
The model I ran used attendance as the Y with a range of variables I believe could potentially have a factor in Brooklyn attendance. For our dummy variables for promotion, I chose to exclude the bobblehead variable. For the day of the week, the dummy variable excluded is Wednesday and for the month of the year, the variable excluded is August. Also, when evaluating team opponents on attendance, the variable I chose to exclude was Vermont. Because our sample size is relatively small with only 42 games, I chose to run two models to avoid potentially including too many variables in one regression model. We face the trade-off of underfitting our model or giving up degrees of freedom. Since there are potential concerns of underfitting our model, we must be tentative in our conclusions in our results and make note of how splitting our model into two may affect it. As for the two models, in the first one, I decided to include the time of the game, day of the week, home winning percentage, temperature, precipitation, and playoff game while in the other I chose to include promotions, opponent, and month of the year. Due to us splitting our variables into two models, omitted variable bias could ultimately be a factor and endogeneity may be prevalent in our results. As stated before, being wary of drawing conclusions based on our results is critical in fully comprehending our results.

Results (regression):

After running the attendance model regression for both our models, the outputs received detailed various factors that impact attendance. It is important to note that variables with a positive coefficient means a positive impact on attendance while a negative coefficient means the variable hurts attendance. For the first attendance model, there were five variables found to be statistically significant. The Friday, Saturday, Sunday dummy variables were positive and statistically significant in comparison to the excluded variable, Wednesday. Temperature was also found to be statistically significant at a p-value less than 0.1 while playoff game had a negative coefficient and was very significant at a p-value less than 0.001. The negative coefficient surprised me as I expected a playoff game would draw in more fans for a game. To expand on this interesting finding, I plotted playoff games on attendance and found that, as seen in the graph, the four playoff games were the lowest attendance games of the year. In our other model which I decided to create to avoid losing degrees of freedom, the output detailed a multitude of variables that ultimately have an impact on attendance. Fireworks and discount were statistically significant at a p-value less than 0.05 meaning these variables have a greater effect on attendance in comparison to the excluded promotion variable, bobblehead. For opponents, we can conclude that playing Lowell and Hudson Valley has a notable effect on attendance in comparison to the excluded variable Vermont. Finally, the model details for us that August and September are statistically significant

First Attendance Model Regression Table

Variable	Estimate	Std. Error	P-value
Intercept/Constant	11345	5704.7	<0.1 .
Home Winning Percentage	-425.6	2557.1	>0.1
Night Game	702.3	625.2	>0.1
Monday	1319.2	924.4	>0.1
Tuesday	1516	961.8	>0.1
Thursday	1333.9	958.2	>0.1
Friday	2025.6	906.1	<0.05*
Saturday	2770.3	976.4	<0.01**
Sunday	2959.5	1096.2	<0.05*
Temperature	-113.9	63.1	<0.1 .
Precipitation	-249.4	376.6	>0.1
Playoff Game	-3642	810.2	<0.001 ***



at a p-value less than 0.1 compared to July. After running these regressions, it is interesting to see that there are only a few statistically significant variables that the Cyclones can actually control. Promotions and day of the week can somewhat be controlled while temperature, opponent, and month cannot. In order to maximize attendance, the Cyclones should look to build an effective schedule with more games on the weekend and more use of fireworks and discount promotions. From our regression analysis, it is important to keep in mind the potential factor of having an underfitted model because of the fact that we split our regression models into two sets of variables. However, because we understood how important it is to not include too many variables in a singular model, we should be confident in the regression results we received.

Second Attendance Model Regression Table

Variable	Estimate	Std. Error	P-value
Intercept/Constant	4997.49	651.82	<0.001 ***
Fireworks	885.94	351.01	<0.05*
Discount	-896.26	373.37	<0.05*
Giveaway	495.07	407.45	>0.1
Themed Night	-202.83	414.81	>0.1
Staten Island	805.32	709.37	>0.1
Aberdeen	-249.81	779.86	>0.1
Lowell	-1906.77	670.72	<0.01**
Mahoning Valley	-977.81	877.62	>0.1
West Virginia	703.07	837.57	>0.1
Tri City	-1053.5	655.51	>0.1
Hudson Valley	-1818.46	661.54	<0.05*
Connecticut	-1141.45	699.66	>0.1
June	84.75	525.01	>0.1
August	912.58	487.76	<0.1 .
September	-1145.01	567.09	<0.1 .

<u>Diagnostics:</u>

To maximize our model performance, we had to run various regression tests. The first thing I did was test for multicollinearity between our variables by using the VIF function in R. If the VIF is below 5 we can be confident in there being no multicollinearity in our model while a VIF greater than 5 can potentially be alarming. A VIF of more than 10 means that multicollinearity is certainly prevalent in our model and adjustments should be made. For the first model, we received no values greater than 5 meaning there are no serious concerns of multicollinearity being present in

First Attendance Model VIF Table

Variable	VIF
Home Winning Percentage	1.79
Night Game	2.15
Monday	3.39
Tuesday	2.91
Thursday	2.47
Friday	3.25
Saturday	2.57
Sunday	4.29
Temperature	2.16
Precipitation	1.27
Playoff Game	1.45

our model. The second model had only one value greater than 5, however, the value itself was just 5.13, which is not large enough to make us uncomfortable with the model we have. The next

Second Attendance Model VIF Table	
Variable	VIF
Fireworks	1.61
Discount	1.83
Giveaway	2.65
Themed Night	2.89
Staten Island	4.17
Aberdeen	4.32
Lowell	5.13
Mahoning Valley	3.46
West Virginia	3.15
Tri City	3.56
Hudson Valley	2.56
Connecticut	2.2
June	2.88
August	3.58
September	2.29

thing I did was a test for heteroskedasticity through the Breusch-Pagan test. In order to do this, I used the "bptest" function in R. For the first model we returned a p-value of 0.7 meaning we fail to reject the null hypothesis for homoscedasticity. In our second model, the p-value received from the BP test was 0.099. This value is above our alpha of 0.05, meaning we also fail to reject the null hypothesis for homoscedasticity. Finally, we tested for autocorrelation in our model. Because our dataset is already created in a time-series element, we do not need to make any adjustments before running the Durbin Watson test for autocorrelation. After running the test, I received a Durbin Watson statistic of 1.739 and a p-value of 0.26 for our first model. The second model's Durbin Watson test returned a D-W statistic of 2.16 and a p-value of 0.338. Because the D-W statistic is near two and our p-value is greater than our alpha value, we can conclude that there is no strong evidence of autocorrelation in either of our models.

Conclusion:

The limitations in the amount of data may have led to various consequences in understanding the true factors that drive attendance for the Brooklyn Cyclones. Using data over the last few years may have helped give us more accurate results however simply by using 2019 data, we were able to get a basic understanding of attendance factors. As the Brooklyn Cyclones look to increase attendance in the years to come, comprehending how each of these variables attributes to putting fans in seats is essential for the club to maximize their attendance.

Data Sources:

https://www.milb.com/brooklyn/schedule/2019/fullseason