Brenden Moore
November 13, 2022

Part A1:

In the first model based on the independent variables given, and running the regression model over the entire range of games at Nationals Park, we get an R Square of 17.55% which is admittedly quite low. For reference, an R Square of 17.55% essentially means that 17.55% of the variation in attendance or our independent variable can be explained by our independent variables.

Each independent variable maintained a P-value below the accepted threshold of .05 allowing all of them to be considered in future versions of the model. However, I made note of the three highest P-values in Holiday, Night, and ATL/PHI none of which appeared in the final model.

Looking towards the coefficients, variables like Opening Day, Garden Gnome/Bobblehead giveaways, and Saturday made the biggest positive impacts while Night was the only variable reflecting a negative impact.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.418892672
R Square	0.175471071
Adjusted R Square	0.168851303
Standard Error	7088.305291
Observations	1131

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	11986435977	1331826220	26.50713147	9.14917E-42
Residual	1121	56323604605	50244071.9		
Total	1130	68310040582			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	26540.20197	444.1803445	59.75095993	0.00%	25668.68351	27411.72042	25668.68351	27411.72042
Opening Day	11815.52337	2004.361899	5.894905196	0.00%	7882.800082	15748.24666	7882.800082	15748.24666
Holiday	4118.834077	1591.408756	2.588168539	0.98%	996.358897	7241.309256	996.358897	7241.309256
Interleague	3302.371246	714.8670887	4.619559773	0.00%	1899.743083	4704.99941	1899.743083	4704.99941
NYY/BOS	12280.47673	2059.271357	5.963505825	0.00%	8240.01656	16320.93691	8240.01656	16320.93691
ATL/PHI	1769.925843	532.4716989	3.323981062	0.09%	725.1724718	2814.679215	725.1724718	2814.679215
Top NL non-division	2502.961034	605.0694987	4.136650484	0.00%	1315.764795	3690.157274	1315.764795	3690.157274
NIGHT	-1584.984441	452.9468468	-3.49927249	0.05%	-2473.7035	-696.265386	-2473.7035	-696.265386
SAT.	4596.7323	574.753955	7.997739311	0.00%	3469.017657	5724.446944	3469.017657	5724.446944
Garden Gnome	4959.500221	927.9709708	5.344456214	0.00%	3138.744672	6780.255769	3138.744672	6780.255769

Part A2:

In trying to improve upon my first model, I first added a variable to separate 2021 and 2022 from before the pandemic with my "Post-Pandemic" variable. I quickly understood that treating 2021 and 2022 equally would have made for inefficient predictions for 2023, as 2021 had attendance restrictions, so I separated with and without restrictions into two independent variables, and decided the 2023 season would still fall into the post pandemic era without restrictions. After these moves, my R Square raised to 30.53% which was a step in the right direction.

Brenden Moore
November 13, 2022

In Regression 4, I attempted to see the impact of individual roster moves by creating a variable for games after the Juan Soto trade or after July 31, 2022. This returned a respectable P-Value but a 6,807.74 coefficient which to predict 2023 is a coefficient that is too high, and the increase post trade could possibly be attributed to individual events such as Soto's return, so I did not include it in future renditions.

Next I limited the window of games included in the model from 2013 onward, as Nationals attendance peaked in the 2013 to 2015 seasons. By changing the window, my R Square improved to 44.62% in Regression 5.

In the next models, I wanted to examine the impact of the month on attendance. Ultimately, I added the summer months (June, July, and August) and the coldest month (April) individually. I considered the two milder months (May and September) equally as they fall inside the school year, but they do not feature the possible extreme weather of April. In the process of adding months, the variables ATL/PHI and Holiday were dropped as their P-Values increased over 5%. After making these adjustments, half of my model's variation could be explained by the independent variables with an R Square of 50.01% in Regression 8. Below is a graph indicating the shifts in attendance over individual months on individual seasons. Including trendlines, it shows that generally, as the year goes on, attendance is increasing.

Next, I unsuccessfully attempted to integrate team performance with the variables Previous Yr Playoffs and Previous Yr Win%. While these slightly improved the R-Square, it greatly shifted my other variables including the key post-pandemic variables. So going forward, I decided not to include these variables, while I now realize the great shift in the coefficients of the other variables was a result of the inclusion of a percentage based variable into a set of binary binary variables.

In Regression 10, I added a column for giveaways which was in addition to the independent variable for garden gnomes/bobbleheads, but this addition resulted in the GG/BBH P-Value soaring to 93.93%, so I rectified this by just having one column representing all giveaways/promotions.

In the next model, I wanted to isolate the impact of individual days of the week, so I created variables for Friday, Saturday, and Sunday while treating Monday through Thursday equally. After doing this, my R-Square rose to 59.11% which would be an amount that the model essentially plateaued at.

In my final models, I added a variable to represent games against the Baltimore Orioles, a match which taps into another fanbase that can attend the game due to the short commute. This had a positive impact of an estimated addition of 4,139 fans.

My final model had an R-Square of 59.70% and included the independent variables, Opening Day, June through August, April, Baltimore, NYY/BOS, Top NL non-division, Friday, Saturday, Sunday, Giveaway/Promotion, Post-Pandemic (no restrictions), and Post-Pandemic (with restrictions).

Brenden Moore
November 13, 2022

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.77267711								
R Square	0.59702992								
Adjusted R Square	0.58679111								
Standard Error	5167.36071								
Observations	566								
ANOVA									
	df	*SS*	*MS*	*F*	*Significance F*				
Regression	14	2.1798E+10	1556984944	58.310512	5.386E-99				
Residual	551	1.4713E+10	26701616.7						
Total	565	3.651E+10							
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	
Intercept	26363.7829	472.504909	55.7957862	0.00%	25435.6516	27291.9142	25435.6516	27291.9142	
Opening Day	13690.357	2206.63304	6.20418381	0.00%	9355.91472	18024.7992	9355.91472	18024.7992	
June	3107.84049	654.851229	4.74587258	0.00%	1821.53018	4394.15081	1821.53018	4394.15081	
July	3167.50693	665.341951	4.76072029	0.00%	1860.58991	4474.42394	1860.58991	4474.42394	
August	2076.10218	639.899759	3.24441783	0.12%	819.160725	3333.04364	819.160725	3333.04364	
April	-3537.0596	682.3171	-5.1838942	0.00%	-4877.3206	-2196.7987	-4877.3206	-2196.7987	
BAL	3926.12913	1296.20795	3.02893463	0.26%	1380.01549	6472.24277	1380.01549	6472.24277	
NYY/BOS	12486.1563	1997.50514	6.25087565	0.00%	8562.49947	16409.813	8562.49947	16409.813	
Top NL non-division	2908.3226	604.942609	4.80760086	0.00%	1720.04673	4096.59848	1720.04673	4096.59848	
FRI.	4987.90103	665.576636	7.49410476	0.00%	3680.52303	6295.27904	3680.52303	6295.27904	
SAT.	7090.12455	612.667116	11.5725561	0.00%	5886.67559	8293.57351	5886.67559	8293.57351	
SUN.	4569.48027	623.39162	7.33003159	0.00%	3344.9654	5793.99515	3344.9654	5793.99515	
Giveaway	1718.03233	503.1716	3.41440639	0.07%	729.663077	2706.40158	729.663077	2706.40158	
Post - Pandemic (no restriction)	-5185.8527	631.319631	-8.214306	0.00%	-6425.9404	-3945.765	-6425.9404	-3945.765	
Post - Pandemic (w/restrictions)	-12633.703	631.468354	-20.006866	0.00%	-13874.082	-11393.323	-13874.082	-11393.323	

Part B:

Using the coefficients of each of my variables, I then made my attendance projections for each individual 2023 game. Each game of the season did get impacted negatively by the Post-Pandemic (no restrictions) variable. Some major series to look forward to with high projected attendances would be the August Boston Red Sox series, San Francisco in July, and the Dodgers in September.

The projected average Nationals Park attendance of 26,412 fans per game is slightly higher than 2022 at 25,017 but is still lower than pre-pandemic levels of 27,899 fans per game in 2019.

AVERAGE PER NATS PARK GAME	26,412

The Nationals pursuit for fans in the seat has been hampered in recent years by an overall decreased interest in the sport, key departures from the franchise such as Max Scherzer, Juan Soto, and Bryce Harper, and the overall recovery from the pandemic.

With a murky future for the franchise including the possible sale of the ballclub and a young roster with little ownership investment on players of high notoriety, it is possible that an adjusted projected attendance considering future variables will be much lower than the 26,412 average fans per game that I projected.

Brenden Moore
November 13, 2022

ATTENDANCE BY MONTH PER YEAR