







# BUSINESS INTELLIGENCE AND ANALYTICS PROJECT

**Goal:** My goal for this analysis is to find the degree to which a team's position in the standings and on-field performance historically impacts their level of home attendance during August and September. The true impact can be seen below as the change in a team's division rankings, win percentage, game-by-game division rankings, and defensive runs above average produces their own respective adjustments to attendance at Target Field.

## Results and its Impact on Fan Attendance:

1. Moving up in the division rankings by one spot is associated with an average increase in the percentage of filled capacity by 2.26%, after accounting for all other explanatory variables, equating to an average increase of about 873 fans.  873 fans
2. A 10% increase in win percentage is associated with an average increase in the percentage of filled capacity by 3.46%, after accounting for all other explanatory variables, equating to an average increase of about 1,367 fans.  1,367 fans
3. Moving up in the division standings by one game is associated with an average increase in the percentage of filled capacity by 0.45%, after accounting for all other explanatory variables, equating to an average increase of about 178 fans.  178 fans
4. A one-run increase in a team's total defensive runs above average (Def) is associated with an average increase in the percentage of filled capacity by 0.06%, after accounting for all other explanatory variables, equating to an average increase of about 24 fans.  24 fans





# ANALYSIS AND WORK PROCESS

**Response Variable:** The average percentage of filled capacity at MLB venues for each month, year, and team specified in the data set (capacity\_percentage)

$H_0$ : The model is not statistically significant for predicting the percentage of filled capacity for each MLB team's venue.

$H_a$ : The model is statistically significant for predicting the percentage of filled capacity for each MLB team's venue.

**Final Model Equation:**  $\widehat{\text{capacity\_percentage}} = 60.9 - 2.26 * \text{Place} + 34.6 * \text{Win\_Pct} - 0.45 * \text{Games\_Back} + 0.06 * \text{Def}$

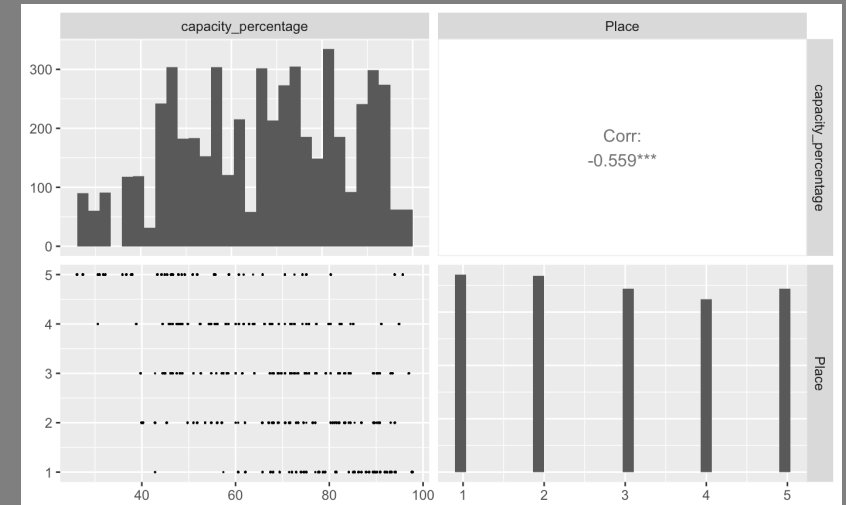
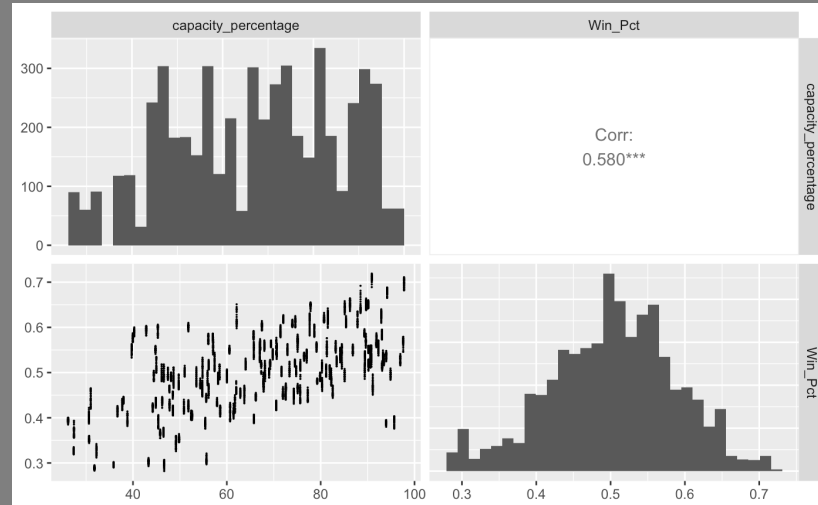
## Data Sources

### Provided:

BaseballGames.csv  
DailyTeamStandings.csv  
Glossary.xlsx  
MLB Venue Capacities.xlsx

### Additional:

FanGraphs season total team batting and pitching statistics from 2017-2019



**Initial Steps:** These two variables, Win Percentage and a team's Place in their division rankings, were highly correlated with our response variable (seen above). Simple Linear Regressions were run for both, individually, highlighting highly significant p-values and impactful slopes, confirming their statistically significant impact on the percentage of filled capacity at MLB venues.





# ANALYSIS AND WORK PROCESS CONT.

**Multicollinearity Issues:** In my process of testing other explanatory variables in search of a more informative model, I found that many offensive statistics coincided with high multicollinearity values when added with win percentage, even though they were highly significant individually. I attempted models without win percentage as well but was not successful in finding a better model fit. The multicollinearity results, based on the model below, highlight one of these tests and its high levels of variation overlap.

$$\widehat{capacity\_percentage} = 92.3 - 1.99 * Place + 90 * Win\_Pct - 0.85 * wRC_+ + 0.897 * fWAR\_batting$$

**Multicollinearity:**

Place	Win_Pct	`wRC+`	fWAR_batting
4.073604	5.403419	6.409114	9.551909

(A threshold of five was used to determine high levels of multicollinearity throughout the analysis.)

**Initial Final Model:**  $\widehat{capacity\_percentage} = 70.5 - 1.88 * Place + 40.1 * Win\_Pct - 0.25 * Games\_Back - 3.67 * FIP + 0.06 * Def$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	70.503068	4.874215	14.464	< 2e-16	***
Place	-1.883292	0.279088	-6.748	1.66e-11	***
Win_Pct	40.057968	5.683852	7.048	2.05e-12	***
Games_Back	-0.248255	0.033459	-7.420	1.36e-13	***
FIP	-3.668847	0.651550	-5.631	1.89e-08	***
Def	0.081258	0.006075	13.377	< 2e-16	***
---					
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Residual standard error: 14.33 on 5242 degrees of freedom  
(153 observations deleted due to missingness)

Multiple R-squared: 0.3972, Adjusted R-squared: 0.3966  
F-statistic: 690.8 on 5 and 5242 DF, p-value: < 2.2e-16

**Caveats:** This model is highly statistically significant, represented by the F-Statistic, overall p-value, and individual p-values, however, our multiple R-squared value is not where we would like it to be for an overall model as it currently sits at a value of 0.3972. Also, there is a bit of a multicollinearity issue in this model (seen on the next slide), however, the multicollinearity is minimal as the win percentage variable is just above the threshold of five. I decided to note this issue but continue with the analysis as multicollinearity will almost always be present to some degree. The impact of the additional variables in this model do more explanatory justice than if I were to remove them due to a slight multicollinearity issue.





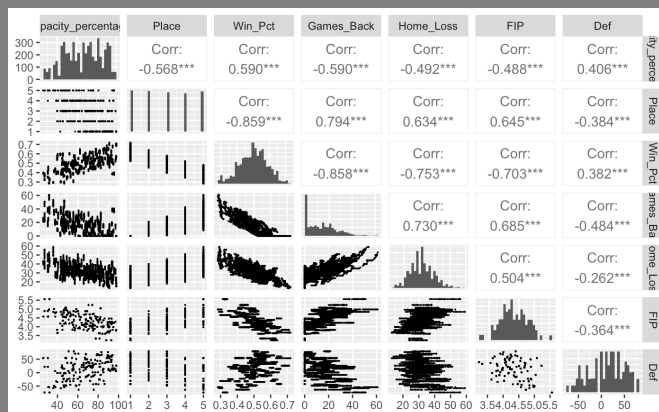
# ASSUMPTIONS

Apart from a minimal aberration in multicollinearity and what seems like a possible outlier issue based on the abnormal points on the equal variance plot and break in the Q-Q plot, all assumptions have been adequately met in the context of our model. All correlations between our explanatory variables and the response variable are significant. Conversely, there does not seem to be any interrelationship in the equal variance plot, and our model's distribution seems to be fairly normal. These results are promising, yet further analysis can be done on this data set to enhance its true statistical impact.

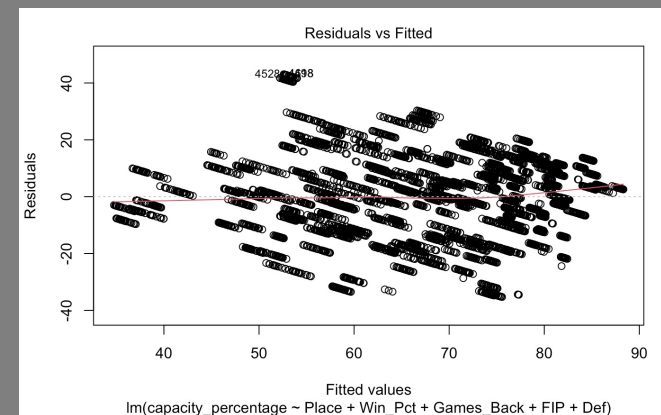
## Multicollinearity:

Place	Win_Pct	Games_Back	FIP	Def
4.031317	5.983194	4.555363	2.102188	1.324239

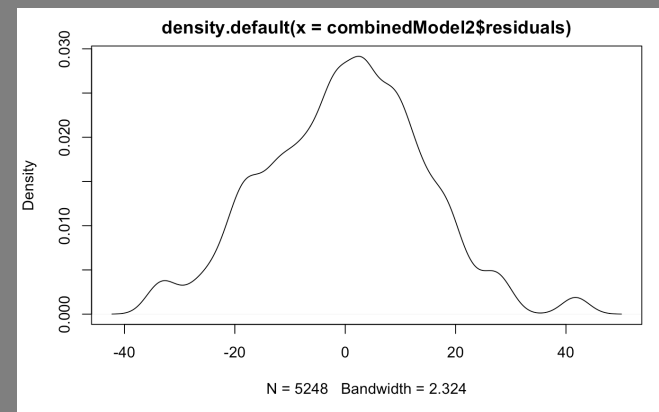
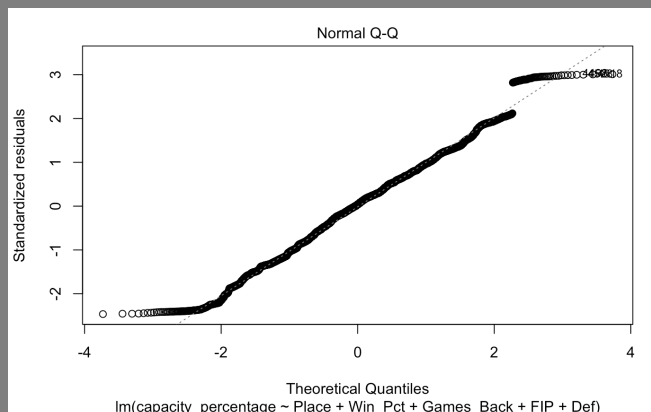
## Linearity



## Equal Variance



## Normality







# OUTLIERS AND MODEL REVISION

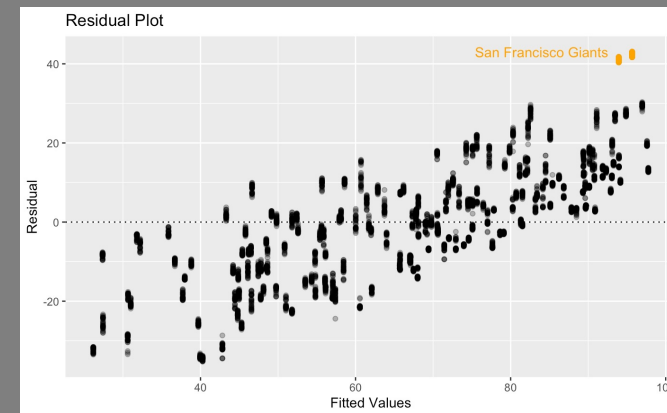
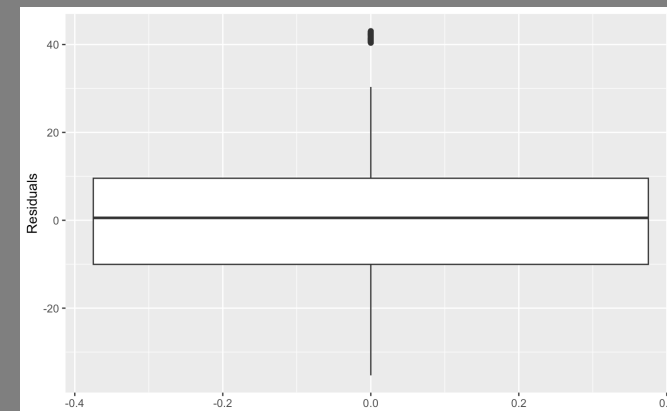
These outliers all interestingly consist of the San Francisco Giants organization. After examination, the Giants' win percentage sits around 39%, yet they still have an average percentage of filled capacity of over 90% at the time. How can this be? There is more to the story for what attracts larger crowds in the MLB, and we know this is true due to our insufficient multiple R-squared value, however, these outliers do significantly deviate from the rest of the data. Due to this, I decided to run this model again, but with these outliers removed. This leads us to our final model (seen below), providing the overall results for the final analysis.

Our multiple R-squared value increased by a fair amount to a total of 0.4502, our multicollinearity issue has subsided a bit, and all our assumptions have been met much more convincingly. This final model provides the interpreted results as seen on the first slide. (The assumption plots have been removed to stay within the limits of the presentation but can be seen in the GitHub if needed.)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.901665   3.388094  17.975 < 2e-16 ***
Place        -2.259787   0.264075  -8.557 < 2e-16 ***
Win_Pct      34.588641   5.269615   6.564 5.76e-11 ***
Games_Back  -0.445601   0.031853 -13.989 < 2e-16 ***
Def          0.062539   0.005796  10.789 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.58 on 5182 degrees of freedom
Multiple R-squared:  0.4502,    Adjusted R-squared:  0.4497
F-statistic: 1061 on 4 and 5182 DF,  p-value: < 2.2e-16
```

## Residual Outliers



(The FIP variable is no longer significant after this revision. Further analysis could be performed to identify the true reasoning behind this issue, but for now, I have removed it from the model.)



## Additional Notes

**Response Variable:** I chose to use the percentage of filled capacity as our model's response variable because using attendance alone would not be informative enough. Each team's attendance is relative; if a smaller stadium had the same attendance as a larger stadium, analyzing solely based on attendance would not provide the correct statistical valuation for both stadium's attendance. Utilizing the percentage of filled capacity as the model's response variable will help with the justification of our answers more than game attendance alone.

**Further Paths to Consider:** Although we have concluded the statistical significance of this model and its rejection of the null hypothesis, the lack of a better multiple R-squared value leads me to believe there is more to be researched on what exactly impacts home attendance. If given more time and data, I would look to research other factors including weather, the time of year of the game based on the location, the number of all-stars on the roster, the number of season ticket holders the organization has, the frequency of theme nights and/or giveaway games, whether current players were close to milestone numbers (like Albert Pujols this past season), the true impact of certain individual players (like Shohei Ohtani), and more. A multiple R-squared value of 0.4502 and the interesting outlier of the San Francisco Giants leads me to assume that there is more reasoning behind home attendance than just how well the team is doing. However, we can conclude that a significant portion of the impact on home attendance is due to on-field performance and the team's place in the standings. Having access to more than three years of data would add more significance to the model, as well. This would allow for the formulation of more significant correlations in fan behavior, therefore allowing the researcher to make more informed inferences moving forward. My next step in this regression would be to analyze the correlations of our model's variables with our response variable to determine if there are any possible data transformations we could perform to find any improvement in predicting each variable's true impact on the percentage of filled capacity at MLB venues. Lastly, if given more time, I would like to analyze this data when compared to early 2000s baseball. 2017 and 2018 specifically were the two most prolific home run-hitting seasons of all time, drawing larger crowds than normal. I would like to research the impact this home run surge has had on MLB attendance as I believe it was a tipping point in the juiced ball era. If I were provided with the opportunity to explore and analyze this topic further, I am confident that I could derive a more significant model explaining the impact in which player performance and other contributing factors mentioned above have on home attendance.

**Data Wrangling and Manipulation:** Gathering the data into a data set with the correct grouping and meaningful information available took some wrangling and manipulation from its initial form in which it was given to me. Mainly, I utilized the tidyverse package's data wrangling verbs in R (`group_by()` with `summarize()`, `mutate()`, `filter()`, `arrange()`, etc.) and joining functions (`inner_join()`, `left_join()`, etc.). With that, I also utilized looping statements and the `as.numeric()` function to manipulate character columns in the dataframes that I needed to be numeric columns. The incorporation of FanGraph's season total statistics for each team was also utilized. The main focuses of the manipulation include the filtration of baseball game data and daily team standing data to August and September only, the separation of each row by its relevant month and year to make it easier to join data sets and interpret the overall data, the creation of our model's response variable, the implementation of teams' season total offensive and defensive statistics through FanGraphs, and the manipulation of in-game statistics to create more informative statistics like residuals of home wins vs home losses or runs scored vs runs allowed. Multiple variables were produced to be possible explanatory variables in future models like offensive totals and residuals. In the final data set used for modeling, I chose to use daily team standings data from August and September (2017-2019) as my base. Attached to that were team-specific statistics based on the month and year in each specific row. This data includes team averages from the specific month and year, totals from the specific month and year, and season totals. (Code can be found in the GitHub.)