OM&IS 324 Group Project - Written Report

Baseball Team Analysis

2014-2015

STATISTICAL TESTING AND VISUALIZATION CREATED IN EXCEL AND TABLEAU

Cameron Thomas, Dakota Gannon, Mark Van Wormer, Kyle Jensen | OM&IS 324 | Fall 2018

Introduction:

For our project we are examining Major League Baseball statistics from 2014-2015 that were sourced from Kaggle.com. We will be investigating correlations between teams wins while considering other variables. The resources we are using to examine the data are Tableau and Excel. This is to assist in visualizing the data using graphs and charts, as well as performing regression analysis on the data set.

Purpose:

The purpose of our project is to compare our dependent variable which is wins and compare them against other variables, such as attendance and runs allowed. We also want to find out if other variables have an effect on wins. Our goal as we examined this dataset ultimately is to figure out what variables have an impact on wins. We will determine our goals by utilizing graphs, statistical measures, regression, and hypothesis testing. Furthermore, we will analyze the various results of the visuals and the statistical analyses to determine our overall recommendations and conclusions on how teams are winning, how to increase team's overall wins, and teams' popularities

Description of the Dataset:

As stated above, our dataset is comprised of multiple stats from major league baseball teams during 2014 and 2015. The dataset we will be analyzing has a total of 48 variables. We found it easier to divide the variables into offensive and defensive statistics. The offensive

variables we thought would have the most potential effect on wins are: hits, 2b, 3b, and homeruns. Hits are described as how often a team hits the ball and gets on base. We are using the abbreviations 2b and 3b to describe doubles and triples, which is how often a team gets a hit and makes it to second and third base respectively. It would make sense that these variables have a strong impact on wins because a team should be more likely to win if they have more opportunities on base, which may lead to more runs scored.

The defensive variables we thought would have the most effect on wins are hits allowed, runs allowed, and earned runs. If a team has a high number of hits allowed and a low number of earned runs, then we believe that the team would have a high number of wins. We also hypothesized that if a team does not allow many hits, walks, and runs, then they would have a better chance of winning more often.

We did an initial analysis of descriptive measures for team wins in each of the years. The table can be seen below. We found that wins are normally distributed with a slight positive skewed and a wide degree of dispersion (kurtosis). Mean, standard deviation, and variance do not seem to differ too much from year to year.

	Mean	Standard Deviation	Variance	Skewness	Kurtosis
2014	81	9.44	89.07	0.02	-1.01
2015	80.97	10.28	105.63	0.04	-0.87

Possible Relationships for Wins:

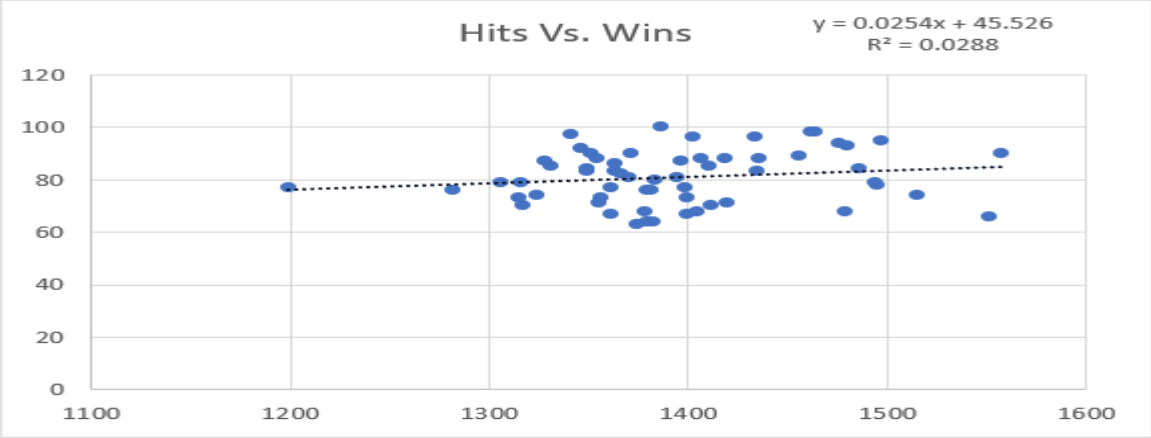
The baseball team data for years 2014 and 2015 have various variables that could potentially correlate with team/franchise wins, such as different types of hits, overall hits, hits

allowed, and many more. However, not all the variables within this dataset fit our overall goals and purposes. For example, the dataset contains repetitive nominal data that are qualitative, such as variables for franchise, team, and retro. Therefore, sticking with variables and attributes that are numerical and discrete will allows us to further our analyses on different impactors of team/franchise wins. In addition, to fit the scope of the project, we will be correlating 5 different variables for wins. We will begin with comparing offensive variables/stats of different teams to see if there is actual statistical evidence that shows the reason why teams are winning. A great variable to begin with would be the total hits of each team. Through hypothesis testing, if there is a linear relationship that appears for hits and wins, then we will further our analysis on different types of hits; such as double hits, triple hits, and homeruns. The next goal for our analysis will be to investigate defensive variables/stats. We will instigate hits allowed and then further our analysis with other defensive variables/stats, such as runs allowed and earned runs. Lastly, we will determine if there is any relation between overall attendance and wins.

Descriptive Statistical Analysis:

Each variable correlation for wins that will be covered will explain and show linear regression hypothesis testing through tools in Excel. Utilizing scatter charts with linear trend lines and the regression/ANOVA tool for R-squared and p-value identification will demonstrate and determine the results of our overall analysis of impactors for wins. Furthermore, we will be assuming that the null hypothesis ($H_0: \beta_1 = 0$) for the linear regression stays true for all variable comparisons for wins, while rejecting the alternative hypothesis ($H_1: \beta_1 \neq 0$).

Hits Vs. Wins:

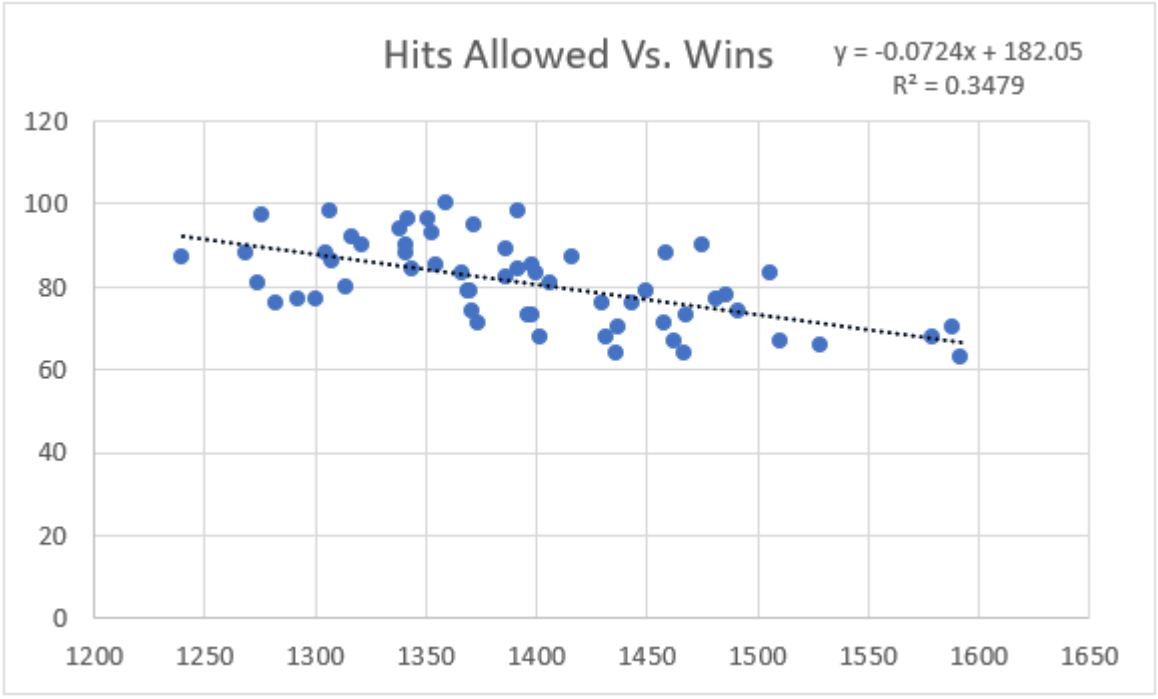


SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.169636				
R Square	0.028776				
Adjusted I	0.012031				
Standard E	9.889826				
Observati	60				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regressio	1	168.0817	168.0817	1.718474	0.195056324
Residual	58	5672.902	97.80865		
Total	59	5840.983			

Analyzing the scatter chart and the regression/ANOVA output shows there to be an extremely weak linear relationship if any at all. The R-squared in the summary output shows to be at .0288, which is extremely low when comparing the 0-1 scale for R-squared. This shows that the goodness-of-fit is extremely poor, with almost no variability of the response data around its mean. Also, an example from the dataset shows that in year 2014, Seattle Mariners had 87 wins and 1328 hits while Tampa Bay Rays had 77 wins and 1361 hits. There are many other teams with similar outcomes and this shows why the goodness-of-fit is poor. This also shows

why offensive variables/stats are a bad predictor and impactor for wins as many other similar variables have the same outcomes. Next, the p-value (Significance F) is surprisingly high. The p-value is at .1951, which is significantly higher than .05. When checking the p-value for the hypothesis testing, 1951 < .05 is not true, therefore the null hypothesis is not rejected. The alternative shows to be not true, and then therefore, there is not a linear relationship between hits and wins. Ultimately, this shows offensive variables/stats of different teams have little to no impact when it comes to wins. This information also shows that there's no reason to further our analysis with other offensive variables/stats, such as double hits, triples hits, and home-runs since the variables/stats have a high variability. For a double hits example, in year 2014, the Baltimore Orioles had 264 double hits and 96 wins while the San Diego Padres had 260 double hits with 74 wins. Triple hits and home-runs also have similar outcomes.

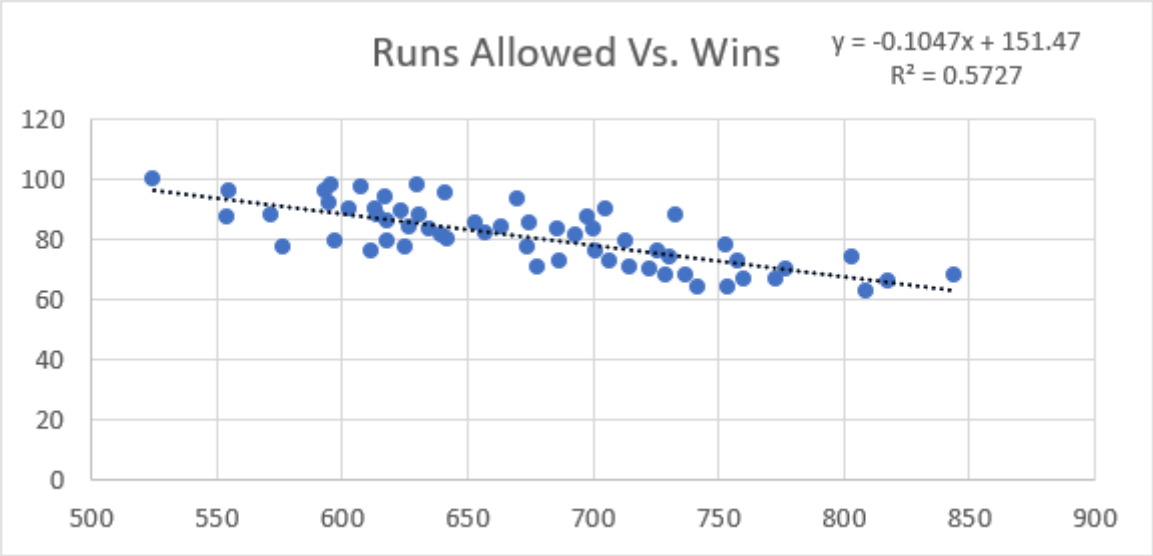
Hits Allowed Vs. Wins:



SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.589795				
R Square	0.347858				
Adjusted I	0.336614				
Standard E	8.104012				
Observati	60				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regressio	1	2031.832	2031.832	30.93768	7.09421E-07
Residual	58	3809.151	65.67501		
Total	59	5840.983			

The first defensive variable/stat shows there to be a negative linear relationship with y = -0.0724x + 182.05. This result makes sense because teams with higher number of wins allow less hits for the opposing team. For example, in year 2015, the St. Louis Cardinals had 100 wins with only 1359 hits allowed. Also, in the same year the Philadelphia Phillies had the highest hits allowed at 1592 with 63 wins. They happen to be the team with the lowest number of wins throughout the MLB. Furthermore, this demonstration has a decent goodness-of-fit since the R-squared is at .3488 and shows that some variability of the response is around its mean. Next, the p-value is shown to be extremely low and is evidently under .05. Therefore, the null hypothesis is rejected and shows that there is an existing linear relationship between these variables. Having an adequate goodness-to-fit and proof of significance of there being a linear relationship shows that hits allowed for teams is a good predictor and impactor for overall wins in the MLB.

Runs Allowed Vs. Wins:

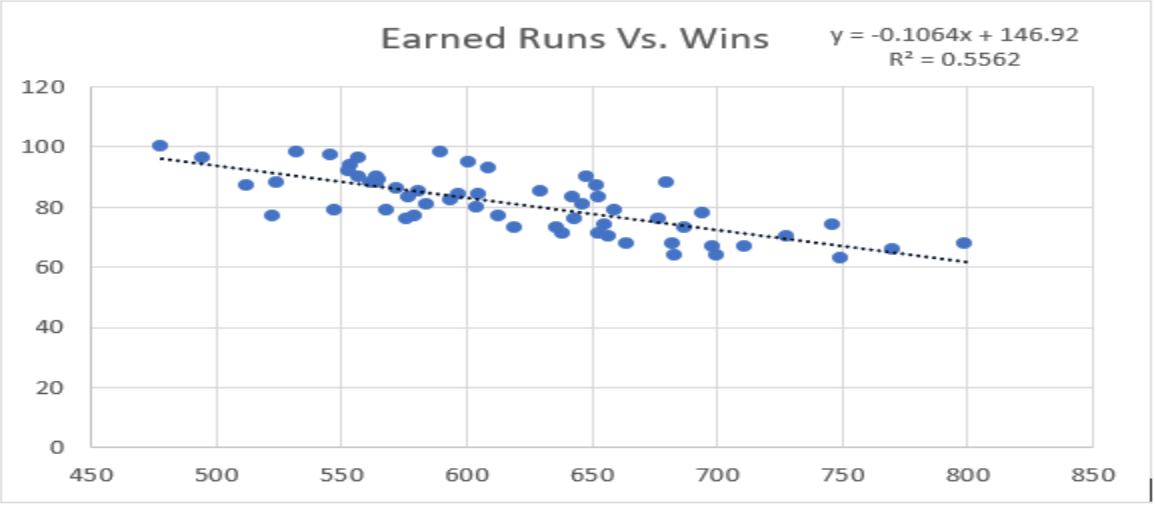


SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.756761				
R Square	0.572687				
Adjusted I	0.565319				
Standard E	6.559974				
Observati	60				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regressio	1	3345.054	3345.054	77.73184	2.66556E-12
Residual	58	2495.929	43.03326		
Total	59	5840.983			

This statistical analysis is similar to Hits Allowed Vs. Wins as it's another negative linear regression (y = -0.1047x + 151.47). A team with lower runs allowed tend to win more baseball games. This regression has another decently high R-squared, which is at .573. The trend line in the scatter chart demonstrates that this statistical test has a good fit with the wins and runs allowed. The p-value in this demonstration is also well below .05. Therefore, the null hypothesis is rejected and shows that there is an existing linear relationship between these variables. Lastly,

the goodness-of-fit and acceptance of the alternative hypothesis shows that this demonstration is a good predictor for wins.

Earned Runs Vs. Wins:



SUMMARY OUTPUT

Regression Statistics	
Multiple F	0.745776
R Square	0.556181
Adjusted I	0.548529
Standard E	6.685469
Observati	60

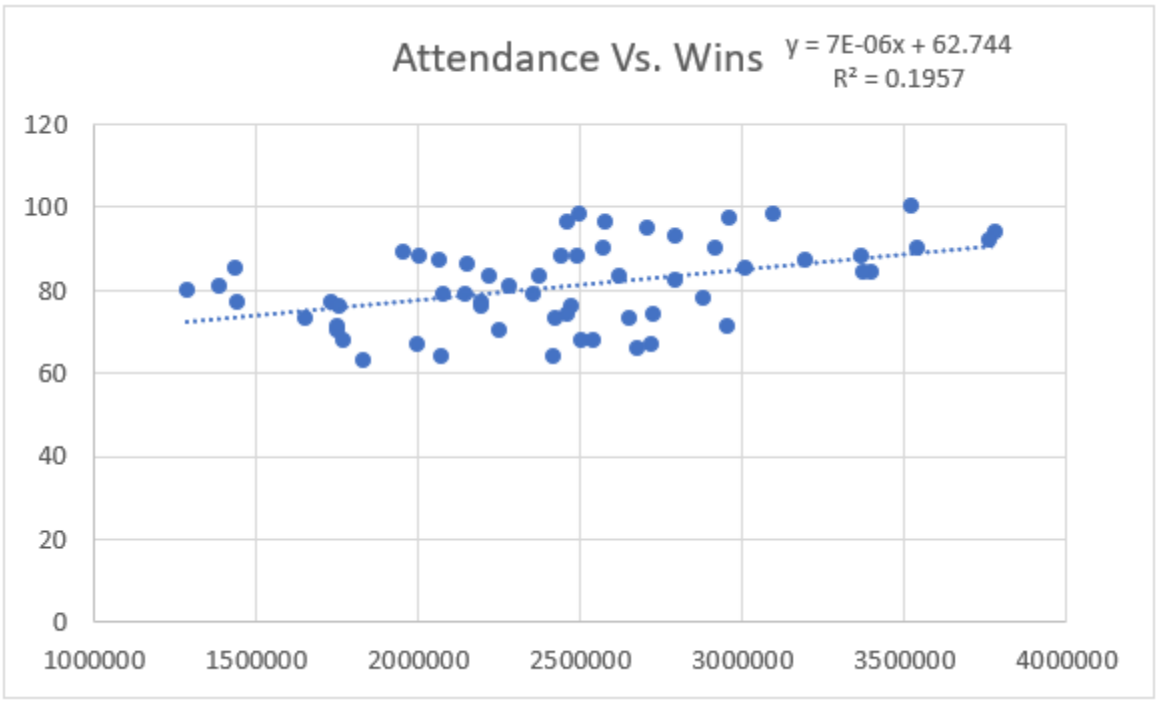
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressio	1	3248.645	3248.645	72.68394	8.1117E-12
Residual	58	2592.339	44.69549		
Total	59	5840.983			

Clarification of what earned runs are: It's any run that scored against a pitcher without the benefit of an error, so therefore it's a defensive variable/stat. Once again, the linear equation is negative (y = -5.2276x + 1043). All defensive variables/stats should be negatively linear due to the dependent and independent variables having an inverse relationship between the two. This last statistical test for defensive variables/stats also shows to have a goodness-of-fit due to the R-

squared being .5562. Ultimately, Runs Allowed Vs. Wins and Earned Runs Vs. Wins show to have a lot of variability of the response data around its mean. Also, the p-value for this test is once again significantly below .05, then therefore we reject the null hypothesis and accept the alternative hypothesis that shows there's a linear relationship between the two variables. Earned runs shows to be a good predictor due to a high R-squared and proven linear regression through hypothesis testing.

Attendance Vs. Wins:



SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.442334				
R Square	0.195659				
Adjusted I	0.181792				
Standard E	9.000134				
Observati	60				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regressio	1	1142.844	1142.844	14.10876	0.000402173
Residual	58	4698.14	81.00241		
Total	59	5840.983			

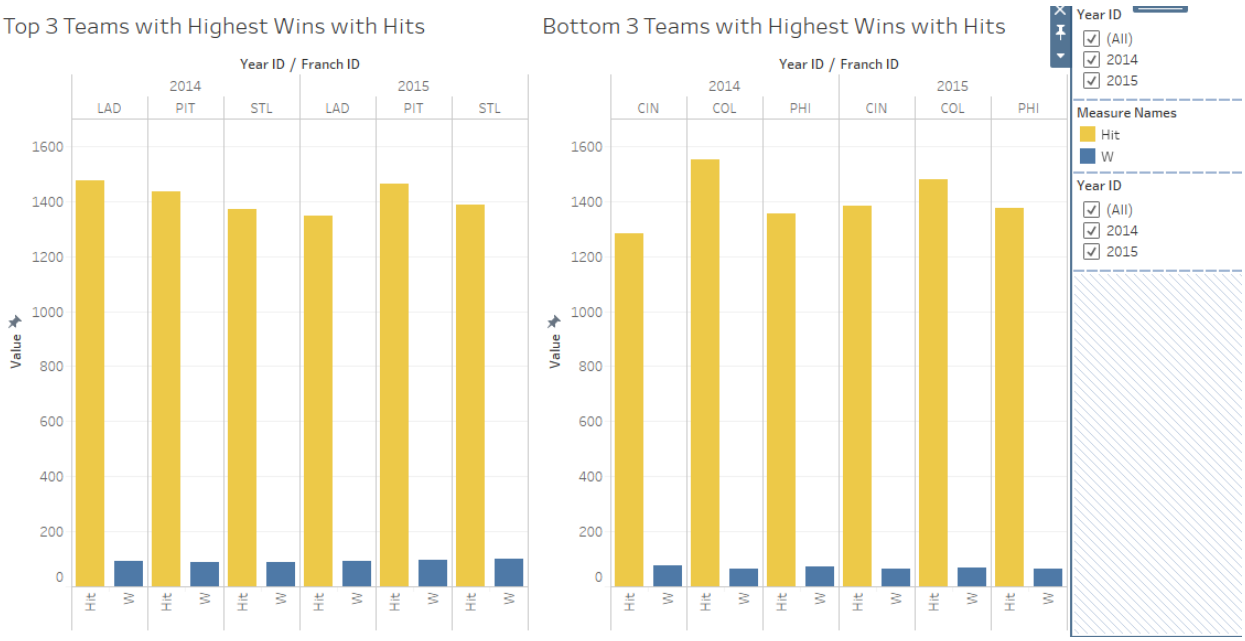
The last statistical analysis that is looked at is Attendance Vs. Wins. The R-squared in this test also seems to be low, which has evidently impacted the goodness-to-fit for the graph. Next, the p-value for this example is significantly low and below .05. The null hypothesis is once again rejected, and the alternative hypothesis is approved and therefore is shown to have significance and a linear relationship. By looking at the evidential information, attendance for a ballpark does increase with higher number of wins in a franchise.

Recommendations:

Baseball teams can be split into two separate categories; offense and defense. For our analysis the focus was on which is more important to earn wins for the team. Does having a weak defense but a strong offense offer a better chance at winning more games, or does a strong defense but a weak offense offer a better opportunity? In this section we are focusing on the offensive stats recommendation, the defensive stats recommendations and then the most important non-game stat for teams, the attendance for the games.

Our analysis found that there is no significant relationship amongst hits and wins. Furthermore, if there was conclusive evidence, then there's an extremely weak goodness-of-fit that shows to be extremely poor with no variability. As mentioned earlier, in 2014, the Rays got 33 more hits during the season over the Mariners (the Mariners had 10 more wins). This shows that recommendations for teams do not concentrate solely on hits as they do not, at least statistically, mean a greater chance of winning. The reason that the other offensive variables vs wins wasn't included was because they didn't have a significant enough impact on the data.

Having a strong offense that includes hits, 1B, 2B, 3B, and runs does not mean a team can always win as there is not a high correlation between them as visualized below. Also, we noticed that the top three and bottom three teams with the highest hits in 2014 and 2015 did not necessarily have the highest or lowest number of wins. For example, STL had the highest number of wins in 2015 but did not have the highest number of hits (1,368). In the same year, CIN having one of the lowest number of wins managed to have more hits than STL. This furthers the strengthening of the fact that a strong offense doesn't necessarily make teams win, rather both a strong defense coupled with a strong offense will better their chances.

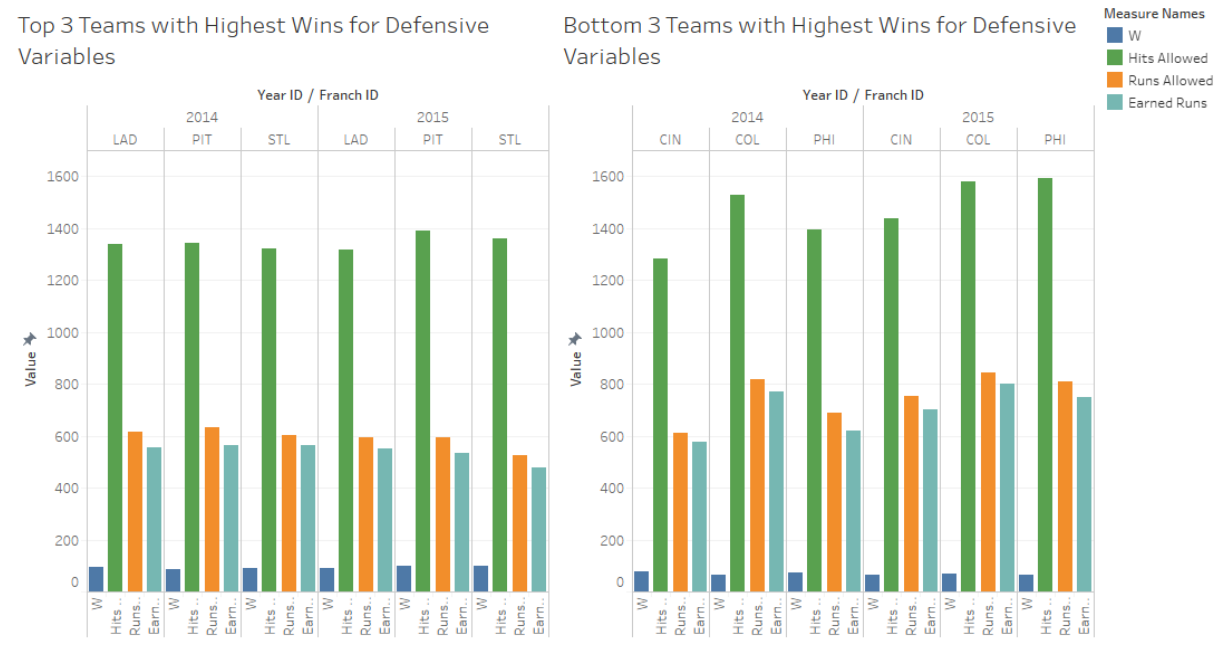


Getting more runs than the other team is how you win in a game of baseball, one method to achieve that is to prevent the other team from scoring. The stats that were analyzed for the defense of the MLB teams include regression analysis of Hits allowed vs Wins, Runs allowed vs Wins, and Runs Earned vs Wins. Comparing the three based from R-squared shows that the strongest is the Runs Allowed vs Wins. This analysis makes sense as the less runs you give up the more likely you are to win, this applies to both teams in the same game though. Some recommendations to make for teams to help prevent Runs allowed would include focusing less on pitching and more on infield play, devoting extra practice time to preventing the number of allowed bases per hit, and to focus on preventing steals and walks.

While the prevention of hits is important to preventing overall scores, the likelihood of a no-hit game based off the data we had would be 0.25%. Therefore, teams shouldn't be focusing on allowing no hits during a game, but rather to prevent the players that do hit from scoring. This can be achieved by focusing less on pitching perfect games and more on the infield and outfield play. The number of runs could potentially be greatly cut down if teams increase the time and

money spent on their training of inner field play, as well as outfield play. With that said, improved pitching is still a great benefactor for increased wins for a team. Another key point to focus on rather than keeping down runs is to prevent the runner from making it past first base. Lastly, the number of hits doesn't include walks, but walks can lead to scores depending on other hits and steals from the team. While a team might not strike out, the number of players aiming to score can be cut down by the prevention of walks.

In reference to the graph below, you can see that the defensive variables (Hits allowed, Runs Allowed, Earned Runs) are incredibly important for increasing wins for MLB teams. The top 3 teams based on wins had on average 200 less hits, allowed as well as 200 less runs throughout the season.



While winning is the most important statistic for the team, revenue is the most important aspect for the owners of the teams. A big contribution towards revenue is brought in by the game

tickets, concessions, and on-site merchandise. The amount of revenue brought in is affected by the number of fans attending the games throughout the season. While most of the research and analysis completed was focused on how teams are more likely to earn wins. There was also regression analysis done on whether attendance affected the number of wins. The R-squared at a mere 0.19 showed that there wasn't a strong linear regression between the two variables. This means that getting more wins throughout the season doesn't guarantee of increased attendance. The in-game statistics do not have a correlation with attendance, so instead, companies should focus more on advertisement, offering discounts or deals, or to expand on the number of seats if they're reaching capacity to ensure more seats are available for fans.

Conclusion:

An initial analysis of the set of MLB data found that Wins were normally distributed with a slight positive skew and wide degree of dispersion, mean, standard deviation, and variance that showed to differ from year to year. A thorough analysis of the data set allowed us to determine impactors for our dependent variable (wins) from the selected independent variables (Hits, Hits Allowed, Runs Allowed, Earned Runs, and Attendance). The analysis showed that the variables that lead towards runs (Hits, 1B, 2B, 3B, Home runs, etc.) were not as statistically correlated to getting wins as the variables that lead to preventing runs (Hits Allowed, Runs Allowed, Earned Runs). For this process, the data was utilized through a linear regression within Excel to find trend lines, R-squared, and p-values. Utilizing these statistical measures allowed linear regression hypothesis testing that demonstrated when either a linear or a non-linear relationship existed. This process allowed analysis and observances of possible correlations between wins and our independent variables. Utilizing our regression models and hypotheses testing allowed

our group to determine that the defensive variables such as hits allowed, runs allowed, and earned runs allowed would have a great effect on wins while discontinuing our statistical analysis on offensive stats due to conclusive evidence for lack of correlation. Since the offensive variables aren't as correlated to wins as the defensive variables, our recommendations leaned more towards defensive options. Teams should focus more on defensive strategies such as earned runs, runs allowed, hits allowed, and prevention since these have more of an impact on their chances of winning. This doesn't mean that teams should completely stop working on hitting and running, but these additional resources shouldn't be their main priority to their current routine. At the very basic level, teams should focus more on preventing runs first and preventing hits second. There are several ways to prevent these, including focusing on infield play, fly ball catches, preventing walks, and limiting the number of bases earned by hits.

Team owners have two primary areas to focus on, their teams getting wins, and making money. We have already covered what to do to achieve more wins, but how can an owner make their team more profitable. A statistic provided by our data set included the yearly attendance for each team. Through analysis we found that the R-squared for attendance vs wins was weak at .19. If a team's winning does not improve the attendance, then what would? We recommended that team owners focus on either improving the amount made during a ballgame by concessions and merchandise sales or to focus on outside factors to draw a larger crowd.