MLB Statistical Analysis

Collin Harrell, Andrew McQueen, Jacob Warner

Data and Decisions

Fall 2021

Executive Summary

With the end of the MLB season finishing just a couple months ago, our group decided to look at five past seasons of stats to analyze for our final project. Baseball is one of the most statistical sports known, thus allowing our group to be able to investigate data anyway we wanted. We decided to pick out a few different team statistics over the five past seasons to see if we can predict whether a team will have a positive record. We decided to use statistics from the 2015-2019 seasons of the MLB to avoid any skewed data after the shortened seasons and many players missing during the COVID-19 epidemic.

As stated, our group wanted to try and predict a team's future record based on past years results. Our dependent variable (DV) to predict is winning percentage or wins divided by losses. We were not interested in if a team made the playoffs based on their record because there would be too many factors comparing a team’s win record to other teams throughout the league. By specifying our dependent variable to winning percentage, we were able to see how successful a team could be during the regular season alone. If the goal was to predict if the team made/had a successful playoff run, then someone could use our model for winning percentage and use it to compare each team’s percentage.

Our data was collected from the website Baseball-Reference (<https://www.baseball-reference.com/leagues/majors/2021.shtml)>. This is a website that tracks each team’s and player’s stats throughout their careers. The website is used by multiple MLB and College teams to track data. Originally, we looked at nine different independent variables (IV) that we felt influenced a team’s winning percentage the most. These include batting average, runs per game, on base percentage, slugging, ERA, runs allowed per game, double plays, errors, and defensive efficiency. Using data across five regular seasons, we had a sample size of 150 to use. Because our data sample was smaller, we needed to make sure our correlation coefficient is very significant, which is discussed further in the modeling steps.

For our process, we first needed to meet the modeling assumptions needed to make the model significant enough to be valid. A quick overview of our steps is first we took the original data set and plotted each of the IV’s vs the DV to see which IVs were more linear. We ended up reducing the amount of IV’s needed to just seven: runs per game, on-base percentage, slugging, earned runs average, batting average, ERA, and defensive efficiency. Then we transformed these new IV’s and ran them through a regression to find which correlated the most and had the most significant impacts on the DV.

After completing our model, one of the major findings we found was that there were noteworthy correlations between multiple independent variables. This was our biggest limitation on trying to predict a winning percentage with those IVs. We had no outliers for our data, making it much easier to use. We then transformed these variables, and, using JMP, we were able to find a worth model. Our model ended up having a standard error of 2.52%, which was low enough to satisfy our assumption. The further steps and analysis of our model can be seen below in the Modeling Steps and Analysis sections. Our final conclusion is that we were successfully able to create a model for predicting a team's winning percentage based on previous years’ data, thus answering our main question: “Are we able to predict how well a team does in the regular season based off their stats?”

Modeling Steps

To produce a model with this data, we need to meet modeling assumptions. Starting with the original dataset, we created scatter plots for each of the seven independent variables against the dependent variable, win percentage (Figure 1). We checked these plots for linearity and found that five of the dependent variables were more linear (Figure 2). The following transformations were used:

* Runs per game: log transformation
* On-base percentage: square transformation
* Slugging percentage: exponential transformation
* Earned run average: square root transformation
* Runs against per game: square root transformation

After these transformations were made, we used Excel to run a standard least squares regression—with the new terms—to check the residual plots and normal probability plot (Figure 3). The residual plots for each of the variables reflected no correlation between residuals. The residuals were evenly distributed and showed no sign of homoscedasticity (Figure 4). The regression also showed the normal probability plot, which was mostly linear, suggesting normality.

We expected that many of our variables would be correlated. For example, an increase in runs per game should come from a combination of increases in batting average, on-base percentage, and slugging. To check for possible assumption violations, we used JMP to run a regression on all higher-order terms and interaction terms. Upon running the regression and removing insignificant terms (with a p-value below an alpha level of 0.05) (Figure 4), we were left with the following seven terms (Figure 3):

* Initial terms:
* Runs per game with a log transformation
* Batting average
* On-base percentage: square transformation
* Runs against per game: square root transformation
* Higher-order terms:
* Batting average: squared
* Natural log of runs per game: squared
* Interaction term:
* Batting average times on-base percentage squared.

Possible linearity assumption violations existed with the new higher-order and interaction terms. We checked this assumption by plotting each term against the win percentage, in Excel. All scatter plots were linear and required no further transformations (Figure 2). Running a regression in Excel produced residual plots to check for homoscedasticity and a normal probability plot to check for normality. The residual plots showed uncorrelated residuals and the normal probability plot was linear. These assumptions being satisfied prompted us to check the remaining regression assumptions for violations.

To check multicollinearity, we used the correlation tool in the Data-Analysis Toolpak (Figure 5). We have a sample size of 150, meaning we needed the absolute value of the correlation coefficient to be less than about 0.15. It was unlikely that our variables would not be correlated. We kept all remaining variables, despite high correlation, (can’t understand how a 1-unit change in IV affects the DV, ie., holding other independent variables constant.)

We used lag variables to check for autocorrelation among our final variables. The lag variables had a sample size of n-1, so that we could compare the second observation to the first, the third to the second, etc. Using the CORREL () function in Excel, we computed each correlation for the terms and their respective lag variables (Figure 6). With a sample size of 149 (150-1), the absolute values of the correlation coefficients needed to be less than around 0.15. The correlation coefficients for each of the variables was below 0.15, with the largest being the natural log of runs per game’s higher-order term at 0.12.

* We ran our final model with (Figure 7) and without the intercept (Figure 8), in JMP, to decide if the error term was unbiased. The intercept needs to stay in the model. At this point, we had verified all of the model assumptions and could run our final model (Figure 9). Our final model contains seven terms, two of which have a p-value slightly above our alpha value of 0.05 (both had p-values of 0.06). These terms could not be removed due to their respective higher-order and interaction terms. Below is our final regression model for win percentage in professional baseball.

,

where:

*Y* = Win percentage

*x1* = On-base percentage

*x2* = Runs against per game

*x3* = Batting average

*x4* = Runs per game

More precise estimates for these parameters are found in Figure 9.

Analysis

To further explain our data, on the “Successful Teams Example” sheet in our Excel (Figure 10), we gathered three teams who have had major success over the past 5 years and plotted their stats. We implemented their raw data on two separate bar graphs for each team. The first graph for each team is their stats that are considered percentages (like batting average) while the other is the whole number stats (like runs per game). This was done to prevent the graph being skewed by percentages & whole numbers. As one can see, the stats for each team are relatively similar, thus the reason why the three teams have been successful.

We also noticed that the runs per game and runs allowed per game should have a close relationship and high significance when compared to whether teams win or lose. We were hoping to see an inverse relationship between the two independent variables themselves to begin. We then decided to look a little further into this relationship to see how the two were related to Win%. This analysis is found on the “RA vs. RG Analysis” sheet in our Excel. We noticed a very linear relationship between the run difference (Runs per Game minus Runs Allowed) and W-L% (Figure 11). We decided to run a regression (Figure 11) to see how the Run difference would predict W/L% as well and listed that below the scatters. We found that the Run difference seemed to be a highly significant predictor of Wins and Losses and were able to display a positively correlated relationship between the two variables.

This model and experiment could be of great use for professional baseball clubs who want to increase their winning percentage. Say a team has a history of struggling, using this formula, they could dissect their stats to try and improve their team. If most of their stats are considered to be near or better than the average, but say their slugging percentage is significantly lower compared to other teams, they would then use this formula to calculate how much better their slugging percentage needs to be in order to increase their odds of winning in the regular season. In a real-world example, the team would then need to try to acquire players who can increase their slugging percentage while still not decreasing their other stats. This method could potentially change the trajectory of a baseball team’s winning record. While our model does work and has a small error, it obviously is not foolproof, as there are many other variables that goes into a winning team in the real world.

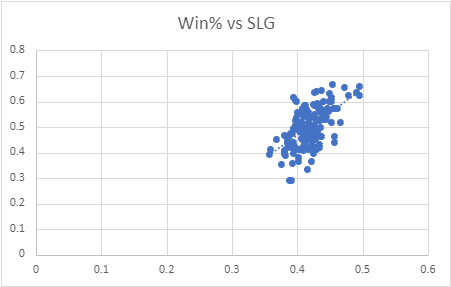
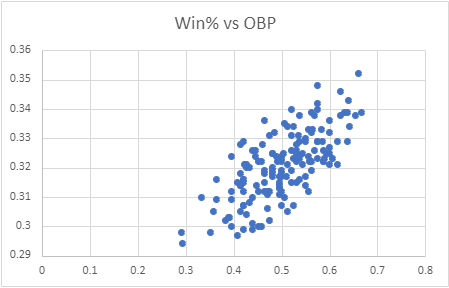
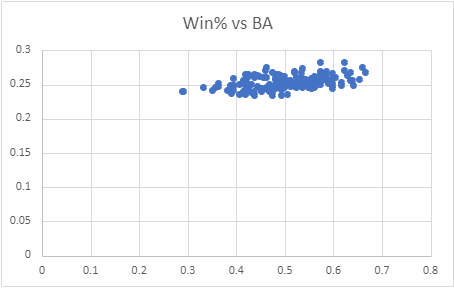
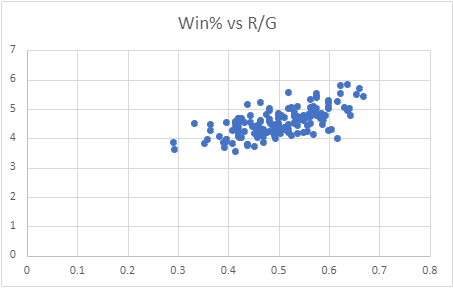
Limitations that we faced in our project mostly stemmed from many of our variables being highly correlated. As one could predict, some variables had very strong correlations when it came to winning ballgames. For example, teams with high batting averages would typically score more runs. Relationships such as this existed between a number of our independent variables, but they were all necessary to test when it came to our final predictions. In the end, referencing the “Final Model” Excel sheet, or Figure 9, we found which variables with their final transformations would give us the best results and ran those in our final simulation model. We found that our variables gave us a highly significant prediction of W-L%, with a root mean square error of 0.0252% and an r-squared value of 0.89. We were extremely satisfied with such a low RSME, and high r-squared value. However, with the multicollinearity assumption being invalid, we struggle to understand the effects of individual independent variables. This is because the coefficients in our final model make up more than just the effects of their respective terms. Overall, the model can predict win percentage very well, but further analysis of individual variables cannot be fully trusted.

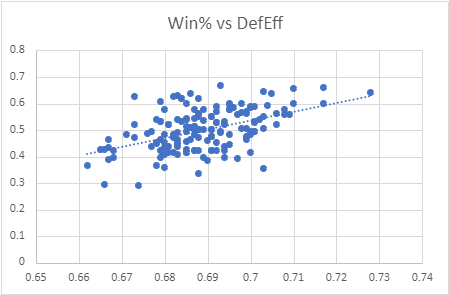
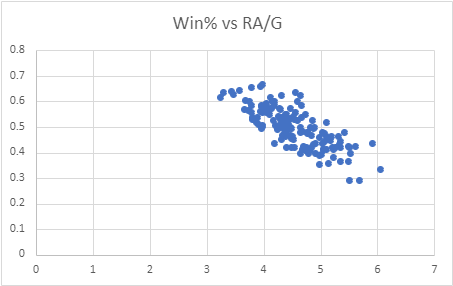
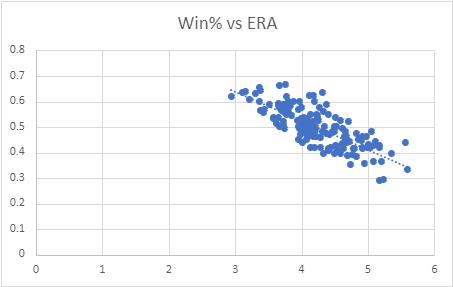
One of our initial plans was to run a simulation for the Chicago Cubs’ future seasons. This would be done by assuming the data was distributed normally. We would find the mean and standard deviation of the Cubs’ respective statistics. We planned to use the Excel function, NORM.INV(), to simulate these statistics many times. Our final model would then use these statistics to predict the win percentages of hypothetical seasons. Since we only used data from five seasons, we did not have a large enough sample size to assume normality for a single team.

From our proposed project plan, our initial question was “Are we able to predict how well a team does in the regular season based off their stats?” As it can be seen from above, we were able to predict a team’s winning percentage based on the seven independent variables selected. With only a small number of limitations, we were able to manipulate our data to produce the formula in our modeling steps. Luckily for us, there were not any deviations to our original plan besides reducing the number of independent variables used and not being able to predict a team’s future seasons since we did not have enough data per team. Our project went smoothly, going exactly how we thought it would go.

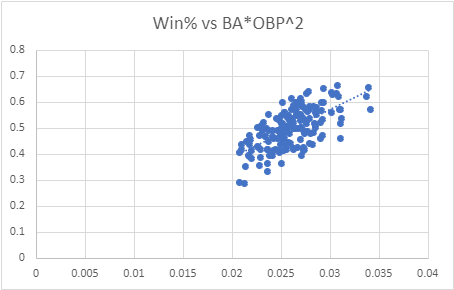
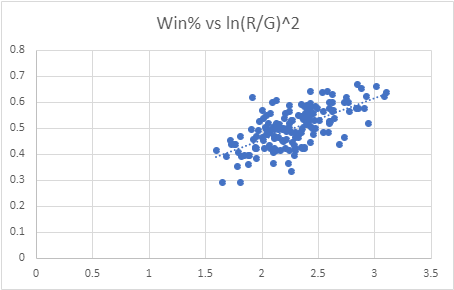
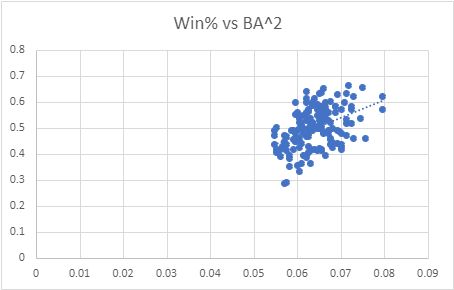
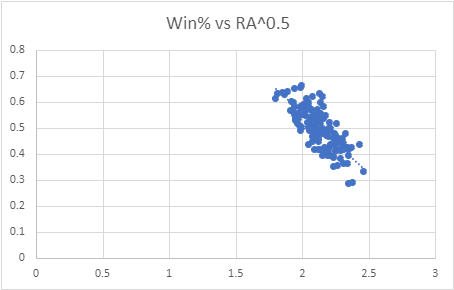
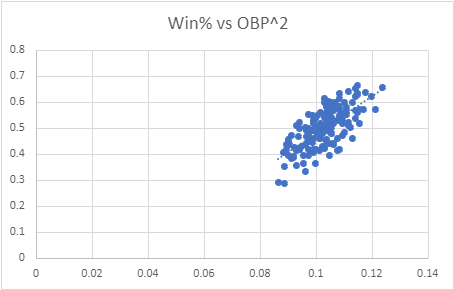
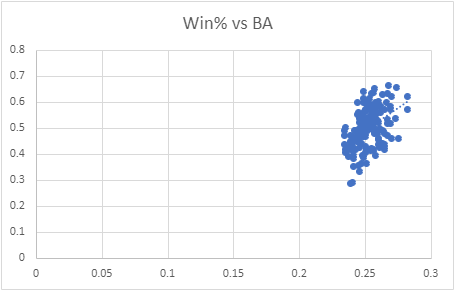
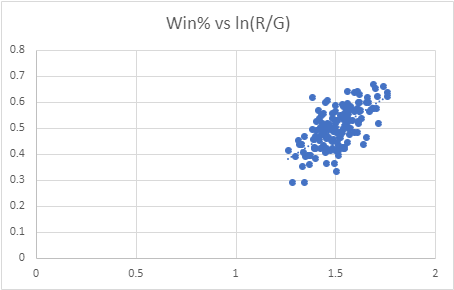
Appendix

* **Figure1:** Original Scatter Plots (IVs vs DV)

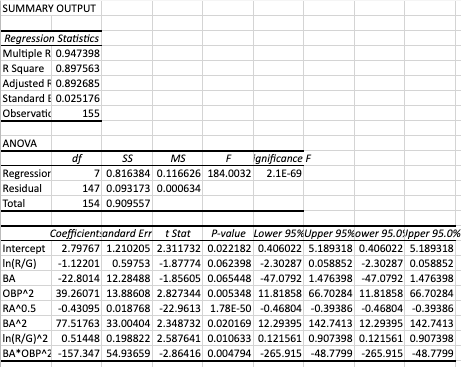
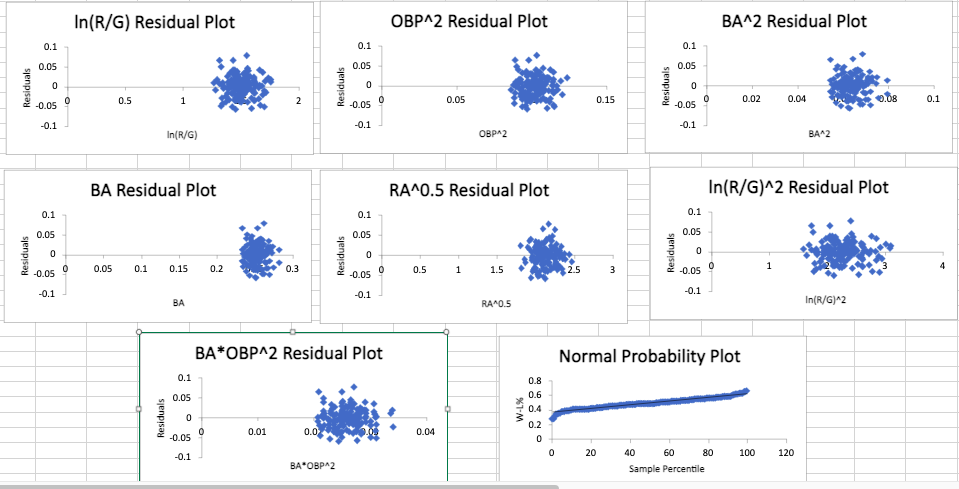




* **Figure 2:** Final Scatter Plots:

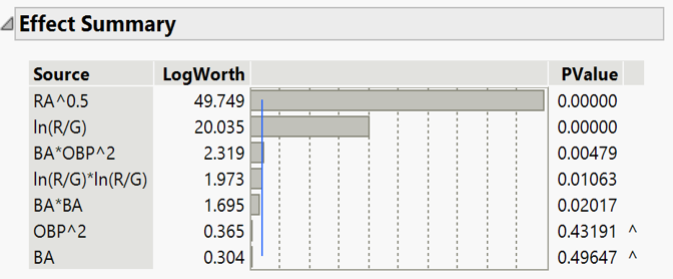
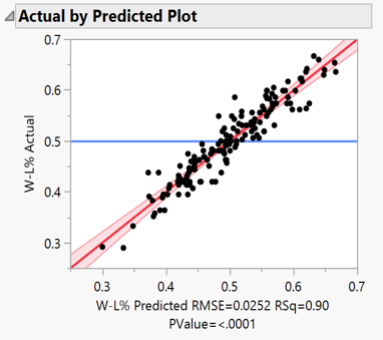


* **Figure 3:** Residuals / Final Terms (Regression Findings + Residual Plots)

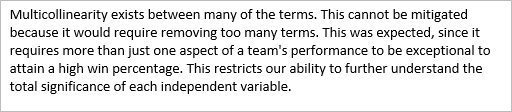


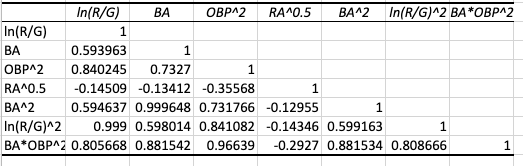
* **Figure 4:** Interaction with HO:



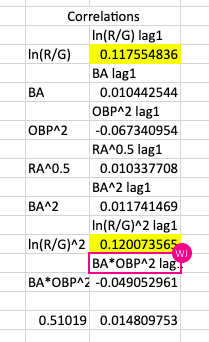


* **Figure 5**: Multicollinearity

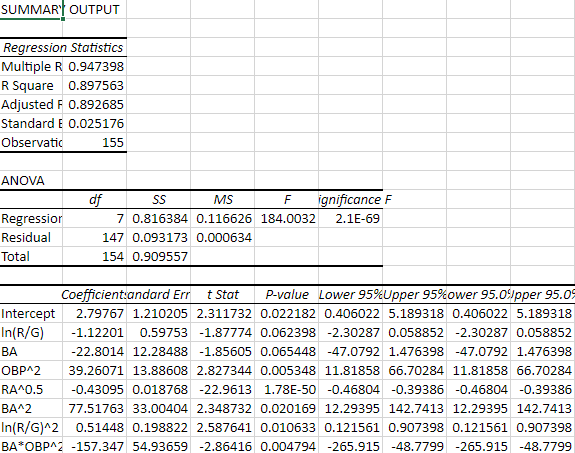




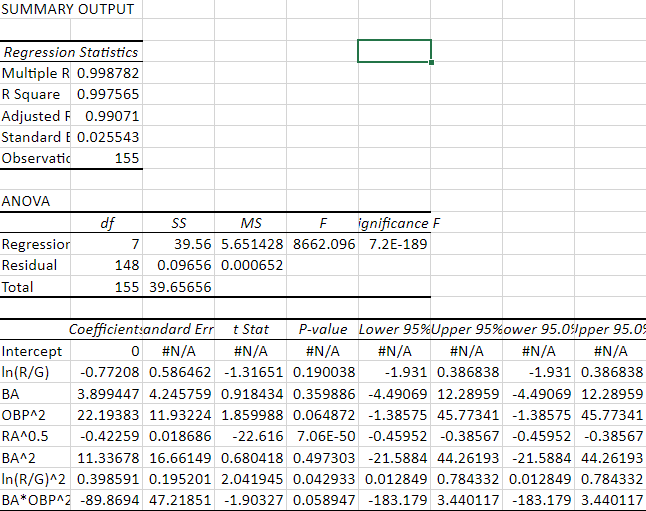
* **Figure 6:** Checking autocorrelation



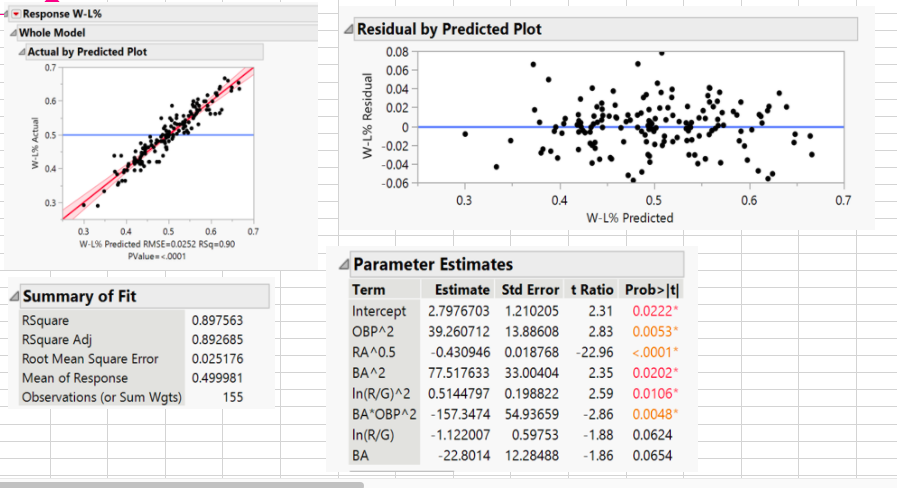
* **Figure 7:** Checking model with intercept



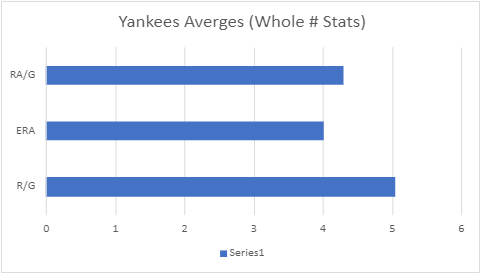
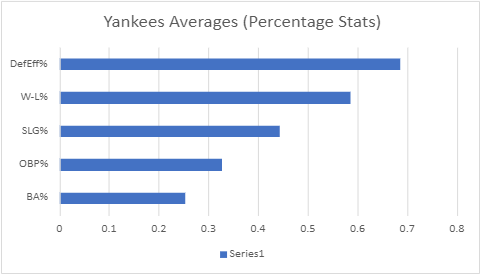
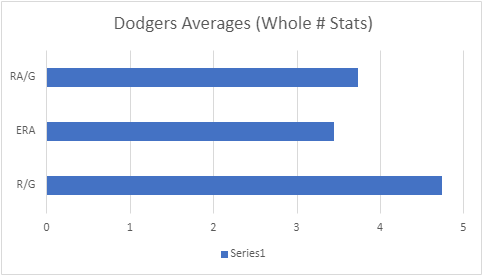
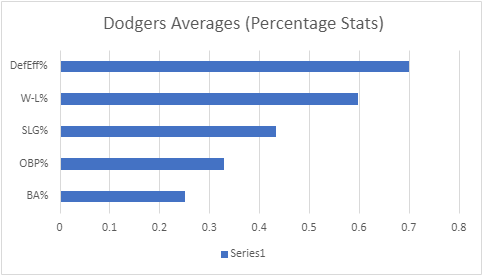
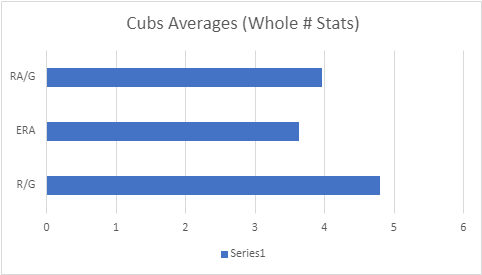
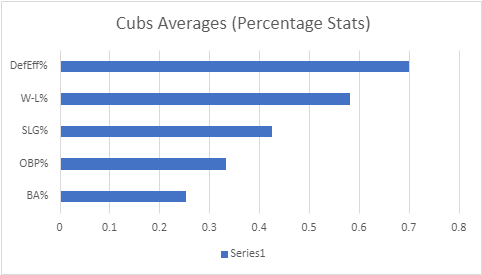
* **Figure 8:** Checking model without intercept



* **Figure 9:** Final JMP Model



* **Figure 10:** Team averages analysis



**Figure 11:** R/G vs RA/G analysis



Chart, scatter chart

Description automatically generated