Multivariate Analysis of MLB Team Batting Statistics

April 12, 2022

BA222 A1

Professor Carlos Casso Dominguez

Jacob Rose U29999110

Dana Winer U13501316

Gabriel Moncau U83571123

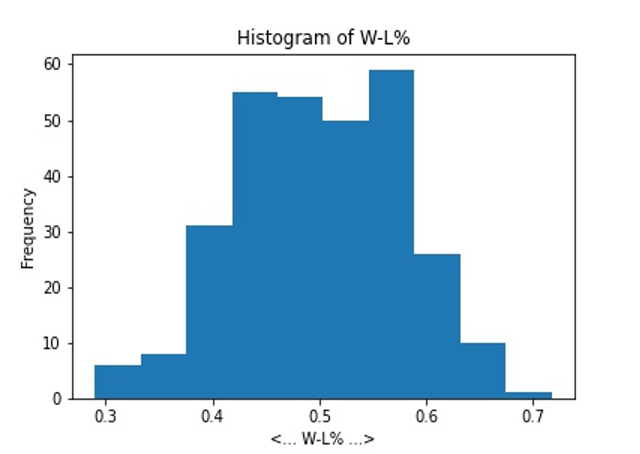
Andres Tamayo U45018174

Introduction

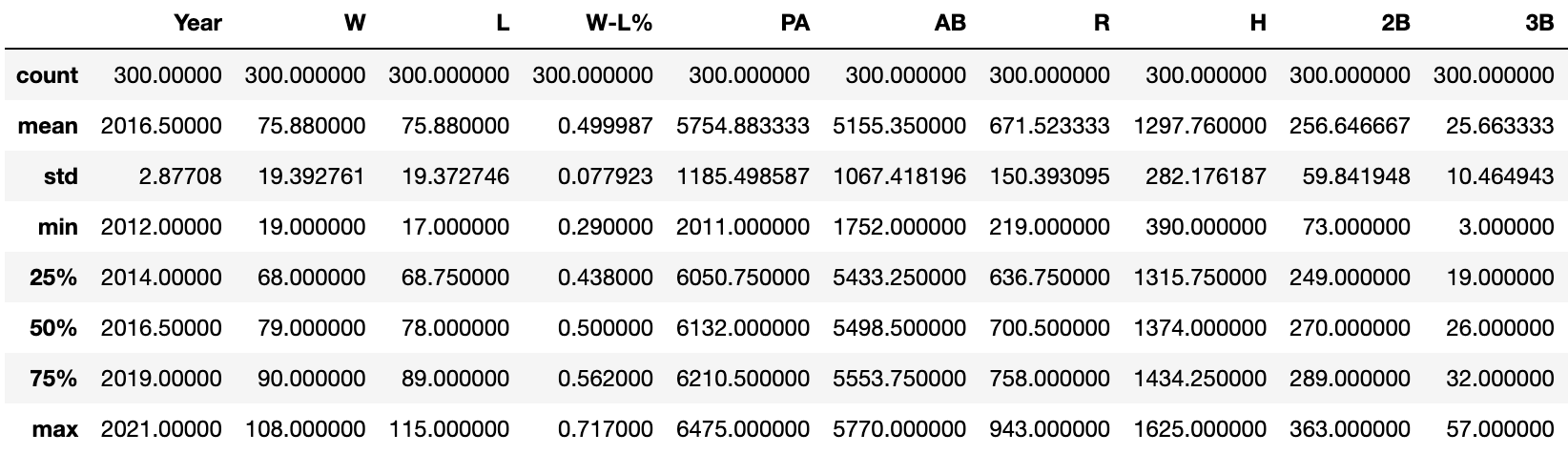
Major League Baseball is a professional baseball organization with the best players in the world. Our motivation for our analysis in team standard batting statistics for the 2012-2021 Major League Baseball seasons is to understand how important different aspects of the game are and how changes across outcomes in the Major League Baseball season can affect team Win-Loss Percentage. Not only does batting average show a hitter's ability to reach base on a swing, but also being at-bat is synonymous with a team being on offense, in which they have the opportunity to score runs to put their team in the lead. The data covers different positions during a Baseball game, which are the following: AB, R, H, Doubles, Triples, HR, RBI, Walks, Strike Outs, Stolen Bases, and Caught Stealing Base. Our group is trying to estimate the relationship between W-L percentage and batting stats such as slugging percentage and on-base percentage. For the controls we are expected to use variables other than OBP that are in the table in the regression analysis and a couple others like SLG, HR, BA, RBI, and BB. These variables will be utilized to reduce bias in our model and create the most accurate estimation of W-L % that we possibly can.

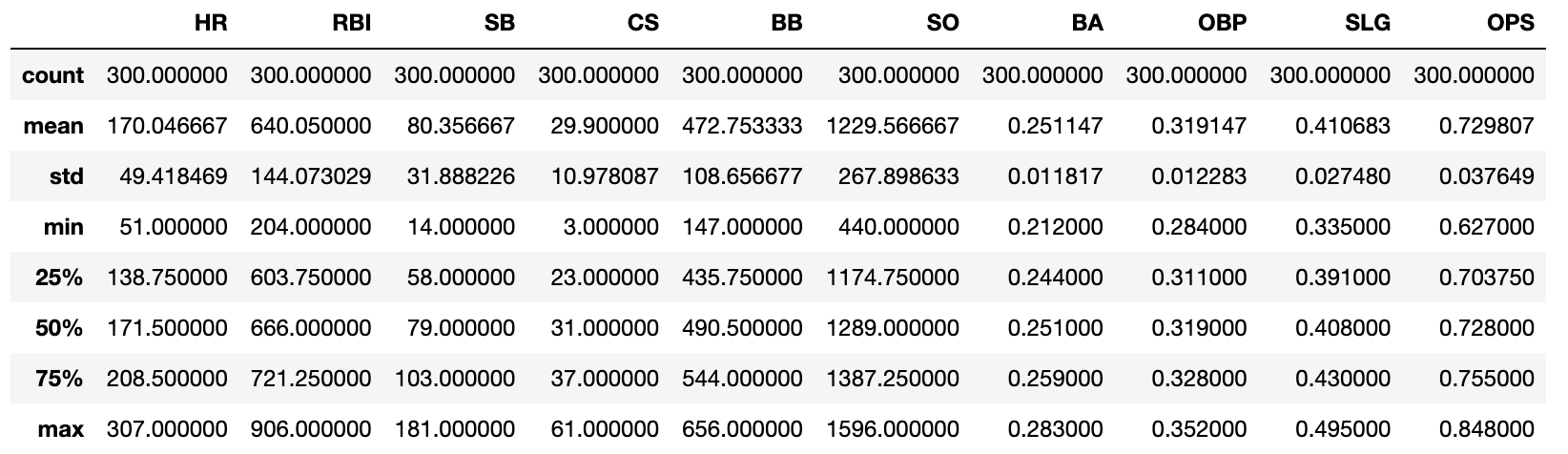
Data Description

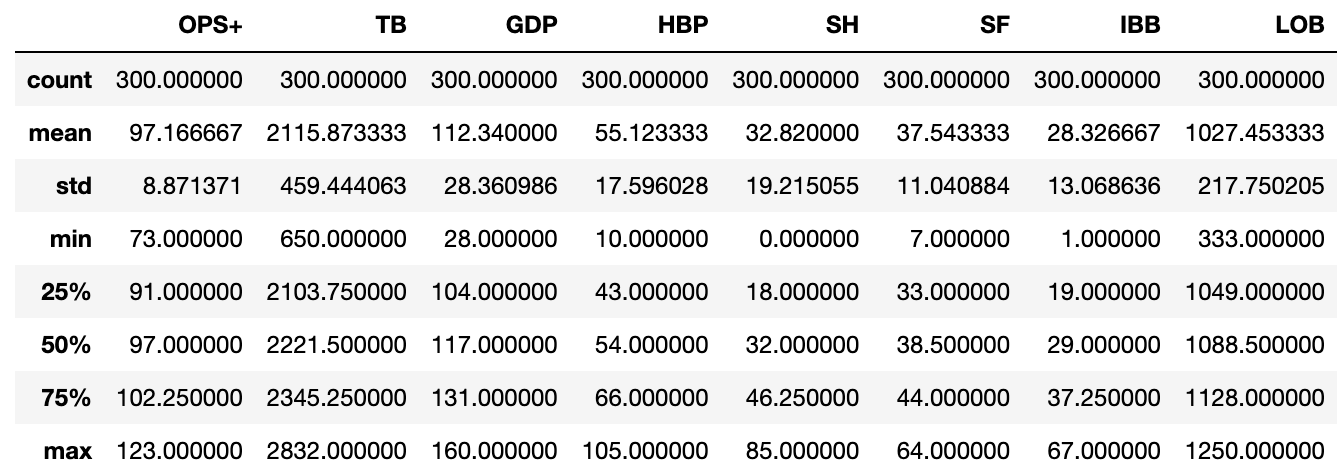
There are 29 data fields in the data set ‘Team Standard Batting Data’. W-L% is the dependent variable in the data set and there are 24 independent variables that are suitable and under consideration to explain the variation in the dependent variable. It is to be noted that W-L% is obtained by dividing the number of wins by number of losses.



The dependent variable W-L% seems to be approximately normally distributed as it is seen from the histogram above. No outliers are visible from the histogram. So, there are no restrictions to assume that the dependent variable is normally distributed although histogram does not show a perfect normal distribution. Descriptive statistics of all the fields in the data set are obtained and shown below and the next page.







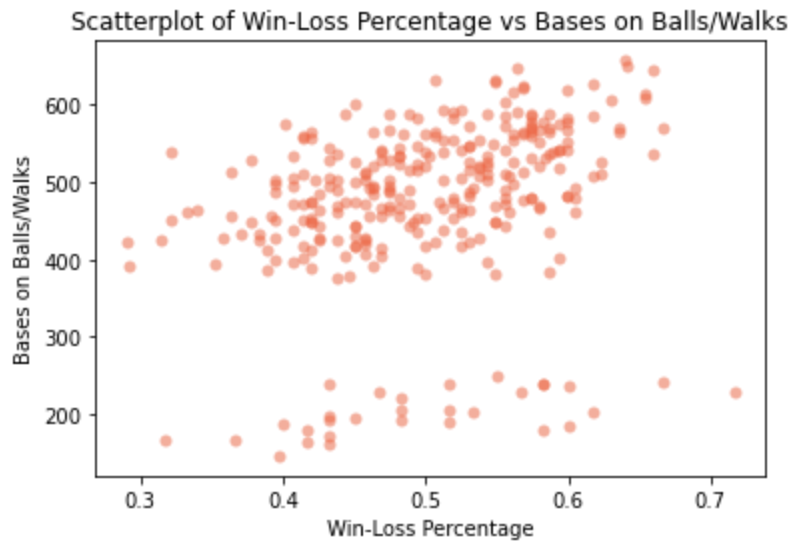
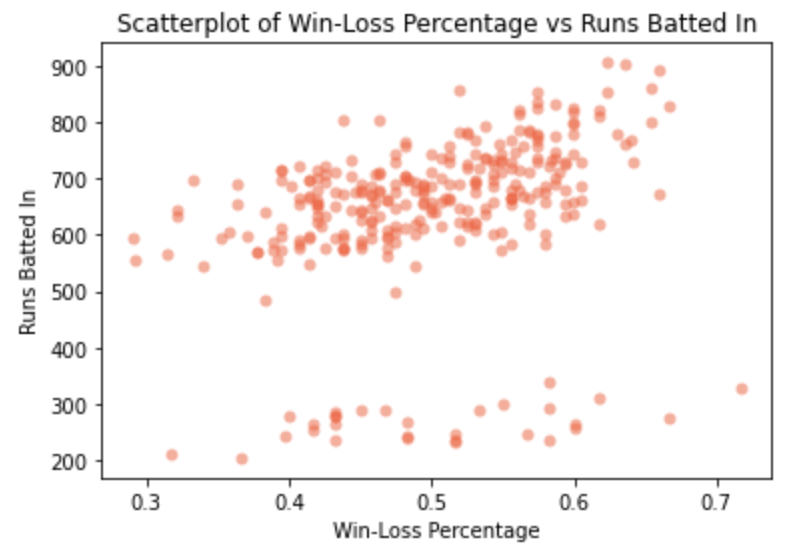
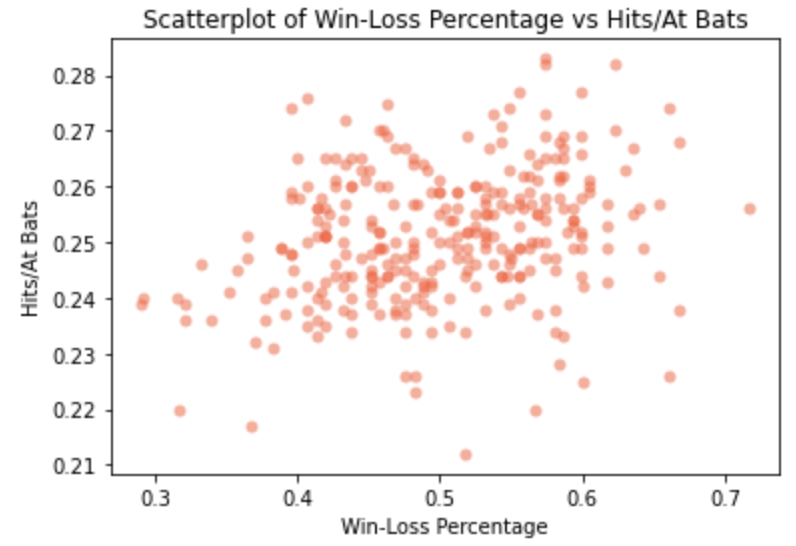
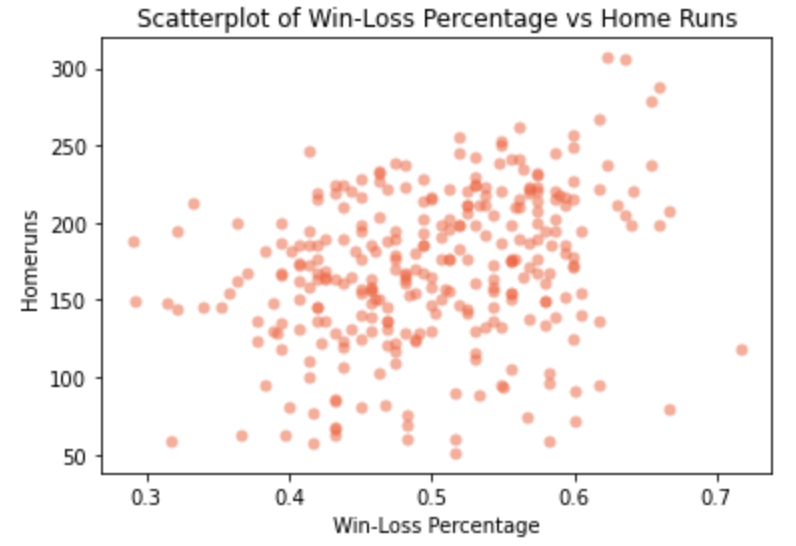
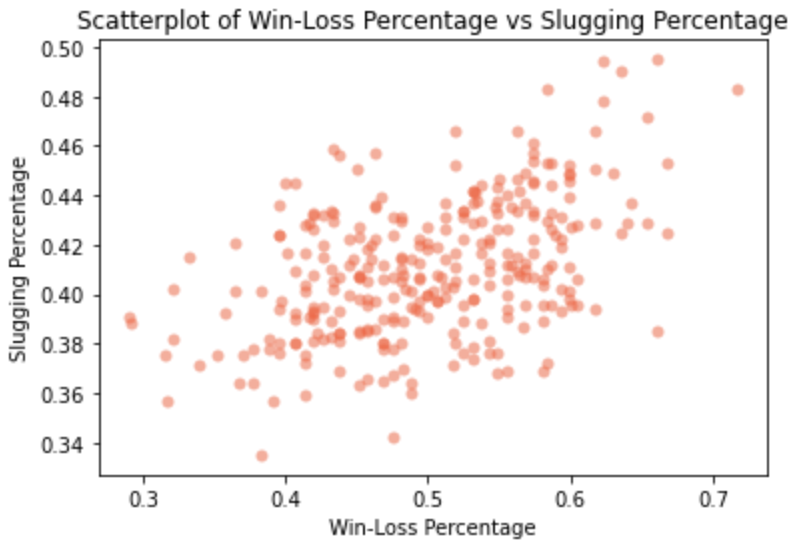
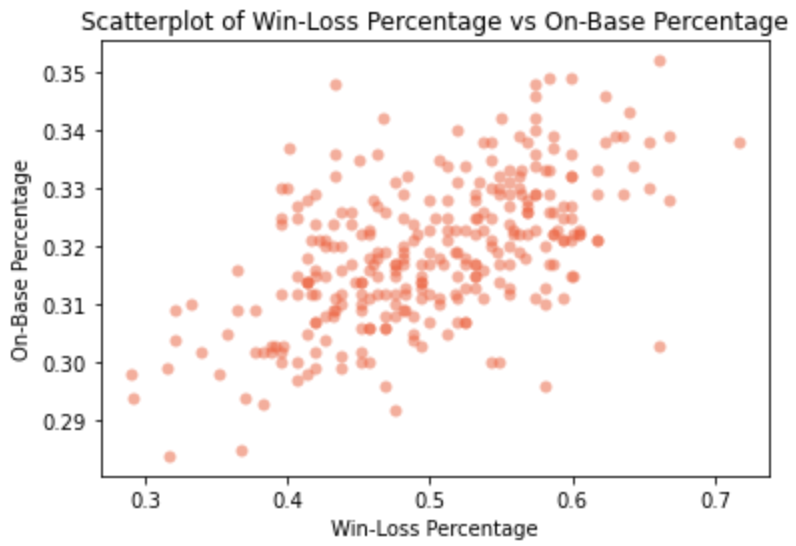
Our summary statistics tell an interesting story for some variables. For example, stolen bases, its mean is 80.36 with a standard deviation of 31.89 which shows a wide distribution of stolen bases throughout the league, as some teams are more aggressive than others.

Separately, for the accuracy of our analysis, we will disregard the variables OPS+ and OPS, as they are both aggregates of other variables in the dataset, with both being products of On-Base Percentage and Slugging Percentage. We will additionally disregard the Wins variable and Losses variable, as they are divided to come to our independent variable of Win-Loss Percentage.

A complete correlation table has been prepared showing correlation among all the variables in the data set. However, the correlation table showing all the possible correlations has not been presented here due to having 32 rows and 32 columns. We are especially interested in correlation between dependent variables and all the independent variables. The table below shows 8 independent variables that show highest correlations with W-L%.

|  | OBP | SLG | HR | BA | RBI | BB | R | IBB |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| W-L% | 0.566623 | 0.460939 | 0.294047 | 0.280768 | 0.274393 | 0.273877 | 0.270729 | 0.191882 |

Below we can see scatter plots illustrating the relation between win-loss percentage and its 6 most correlated independent variables. Each of these graphs shows a positive relationship between the independent variable and Win-Loss Percentage. We can see some separation in data as there was a shortened season due to COVID in 2021, cutting games from 162 to only 60.



As already mentioned, our dependent variable is W-L%, which we will be evaluating using six independent variables: OBP, SLG, HR, BA, RBI, and BB. One or two independent variables can be used as control variables in the regression model to see the effect of another independent variable on the dependent variable, W-L%. For example, we can control RBI and see the effect of other independent variables on the dependent variable.

Regression Analysis

We prepared for our regression analysis by determining which of our possible independent variables would be most suited to being the main independent variable.



By analyzing the correlation matrix between W-L % and the other variables, we decided to use OPB, or On-Base Percentage, due to its relatively high correlation value of 0.567. OBP is defined as how frequently a batter reaches base per plate appearance. Times on base include hits, walks and hit-by-pitches, but do not include errors, times reached on a fielder's choice, sacrifice bunts, or a dropped third strike. From that univariate regression, we then added any relevant control variables sequentially to get more accurate multivariate models. Based on each control’s p-value/t-stat and effect on the r-squared we chose whether to keep them or remove them from our model so as to minimize multicollinearity. Finally, from that process we were able to generate the data shown in the table below which displays our univariate regression followed by each of our multivariate regressions.

|  | W I | W II | W III | W IV | W V |
| --- | --- | --- | --- | --- | --- |
| Intercept | -69.3361\*\* | -19.1367 | -31.5173 | 9.6490 | 15.0224 |
|  | (27.9701) | (21.4828) | (21.1634) | (14.9303) | (14.6422) |
| OBP | 455.0137\*\*\* | 161.2900\*\* | -108.5486 | 206.6882\*\*\* | 170.3813\*\* |
|  | (87.5757) | (69.3041) | (94.7350) | (68.3718) | (67.3961) |
| HR |  | 0.2561\*\*\* | 0.2541\*\*\* | -0.1031\*\*\* | -0.0553\*\* |
|  |  | (0.0172) | (0.0168) | (0.0233) | (0.0258) |
| BA |  |  | 393.5442\*\*\* | -314.1793\*\*\* | -285.4938\*\*\* |
|  |  |  | (96.8526) | (78.4121) | (76.9102) |
| RBI |  |  |  | 0.1511\*\*\* | 0.1273\*\*\* |
|  |  |  |  | (0.0085) | (0.0103) |
| IBB |  |  |  |  | 0.2165\*\*\* |
|  |  |  |  |  | (0.0551) |
| R-squared | 0.0831 | 0.4742 | 0.5020 | 0.7590 | 0.7710 |
| R-squared Adj. | 0.0800 | 0.4707 | 0.4969 | 0.7557 | 0.7671 |

Each variable’s coefficient for the respective model is displayed with a number of stars (\*) next to it equal to its significance (\* = significant at 0.1, \*\* = significant at 0.05, \*\*\* = significant at 0.001).

By adding controls to our analysis we were able to raise the r-squared from 0.0831 in our univariate model to 0.7710 in our final multivariate model, showing an increase in accounted-for variance of 0.6879, or 68.79%. This also fundamentally changed our understanding of the relationship between W-L % and On-Base Percentage in this dataset by greatly decreasing the perceived influence of On-Base Percentage in win rate. It also, crucially, showed the importance of Home Runs in Win-Loss Percentage, which alone increased our r-squared by 0.3911. From our final model we are able to see that, in this dataset, HR, BA, RBI, and IBB are all heavily related to win-rate, with OBP also being an important factor. It also tells us that when all the controls above are 0, win rate will on average be 15.022%, as well as how much Win-Loss Percentage will change by on average when any of the given variables changes by one unit (shown by their respective coefficients).

Conclusion

In conclusion, our multivariate regression drastically improved our model and we were able to come up with an extremely accurate model to predict win rate based on a number of factors. Our final model is as follows:

W-L % = 15.022426 + [OBP \* 170.381289] + [HR \* (-0.055263)] + [BA \* (-285.493848)] + [RBI \* 0.127252] + [IBB \* 0.216458]

By utilizing dependence techniques, looking at cause-and-effect relationships between the variables to come to our final model conclusion. This brings us to the conclusion that a combination of On-Base Percentage, Home Runs, Hits per At Bat, Runs Batted In, and Intentional Walks generate the most accurate estimation of Win/Loss Percentage for MLB teams over the past 9 years. By adding these variables, we are reducing omitted variable bias slowly through the model, however, there is a possibility that we are missing variables that could improve our model and therefore succumbing to omitted variable bias at the same time.

Concerning future implications of this model, it shows the importance of certain variables over others. Seeing the negative slope on the Home Run variable versus the positive slope on the Runs Batted In and OBP leads me to the conclusion that having consistent hitters should take priority over power hitters when the main goal of every MLB team is to increase Win-Loss Percentage (with exceptions). This, of course, is contingent that the model is proven as causal in the future through more in-depth analysis. The main goal of this analysis was to understand how important different aspects of the game are and how changes across outcomes in the Major League Baseball season can affect team Win-Loss Percentage. I believe we have achieved this by closely speculating which variables greatest affect the estimation of Win-Loss Percentage over the course of an MLB season.