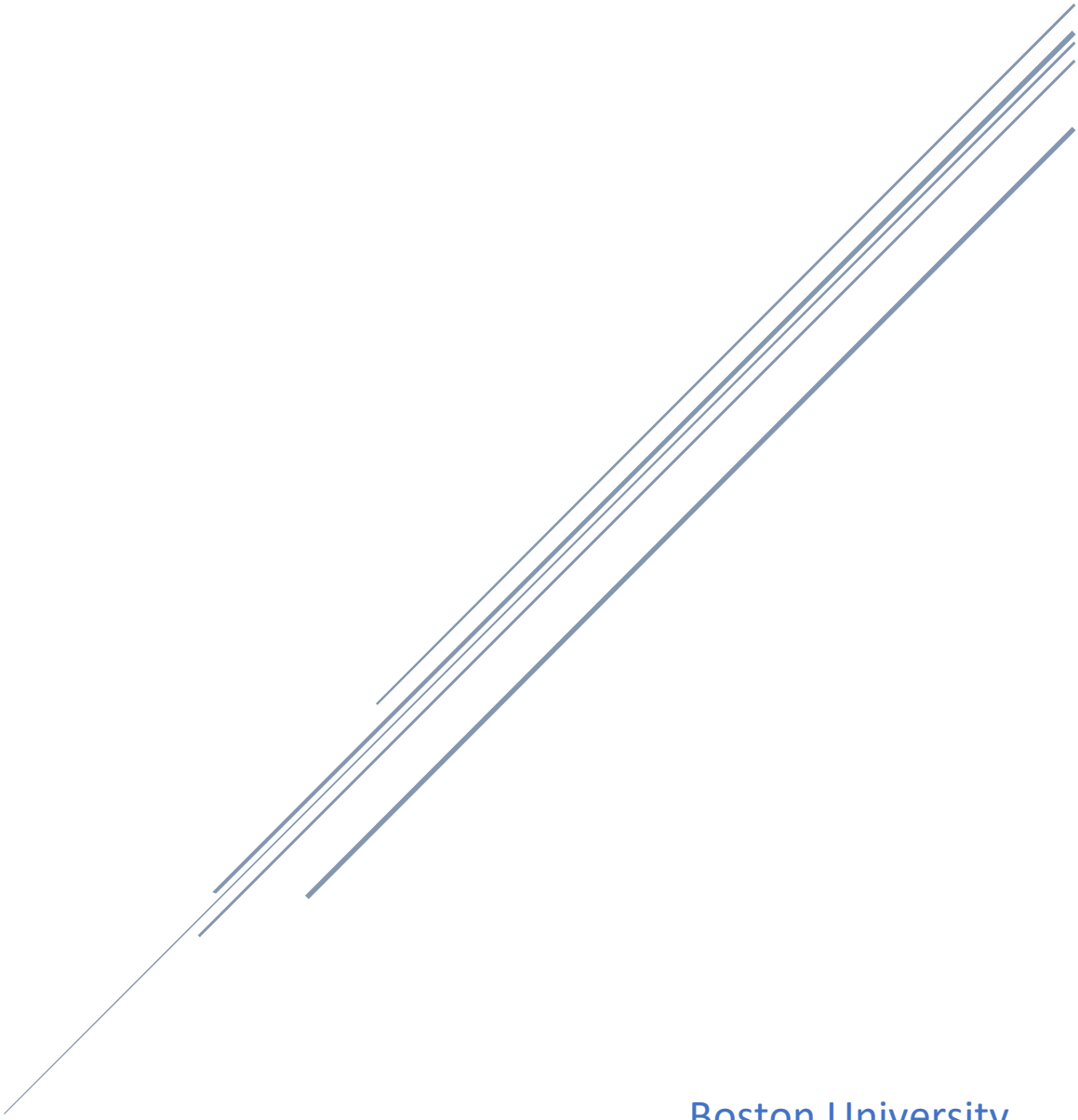
ANALYZING BASEBALL STATISTICS ACROSS CULTURES

: A STUDY OF THE KBO



Boston University
Junho Eum

Table of Contents

Abstract

The Korean Baseball League (KBO) was established in 1982, significantly later than other professional baseball leagues such as Nippon Professional Baseball (NPB) in Japan and Major League Baseball (MLB) in the United States which were founded in 1936 and 1876, respectively. Consequently, the KBO has been influenced by both the Japanese and American baseball cultures, resulting in the development of its unique playing style. This research paper aims to analyze and compare the playing style of the KBO with that of the MLB and NPB by studying their respective statistical data. The research question for this study is: "How does the KBO differ in its statistical characteristics compared to the MLB and NPB, and how do these differences contribute to its distinct playing style?"

Introduction

A myriad of studies has explored various aspects of baseball statistics, such as player performance, game strategies, and team dynamics, in different leagues. Nevertheless, limited scholarly attention has been given to comparative analysis across different professional baseball leagues, specifically between the KBO, MLB, and NPB. This study seeks to fill this gap in the literature by conducting a cross-cultural analysis of baseball statistics through the lens of Pythagorean Expectation, a formula used to predict a team's win rate based on runs scored and runs allowed. Bill James, the pioneer of sabermetrics, initially developed the Pythagorean Expectation equation:

Pythagorean Expectation=(RS^2)/(RS^2+RA^2)

```

To conduct a comparative analysis, this study will analyze pitching and batting datasets obtained from Kaggle, employing methods of preprocessing and parameter selection to build a model based on the Pythagorean Expectation formula.

## Methodology

### Preprocessing & parameter selection

The first step involved importing two datasets (pitching and batting data) and checking for null values in each column. This revealed missing data points from the past due to the absence of a
```

stat recording system at that time. As the research focuses on recent advancements and differences in playstyle in the KBO, duplicates and rows with null values were removed.

To identify the variables with the highest correlation to the dependent variable, win-loss percentage, a correlation heatmap was generated (Figure 1-1). Subsequently, the heatmap was focused on the correlation between win-loss percentage and other variables (Figure 1-2). This plot highlighted the independent variable with the highest correlation to win-loss percentage, while excluding variables like wins, losses, and run average, which exhibited high collinearity with the response variable 'win-loss percentage'.

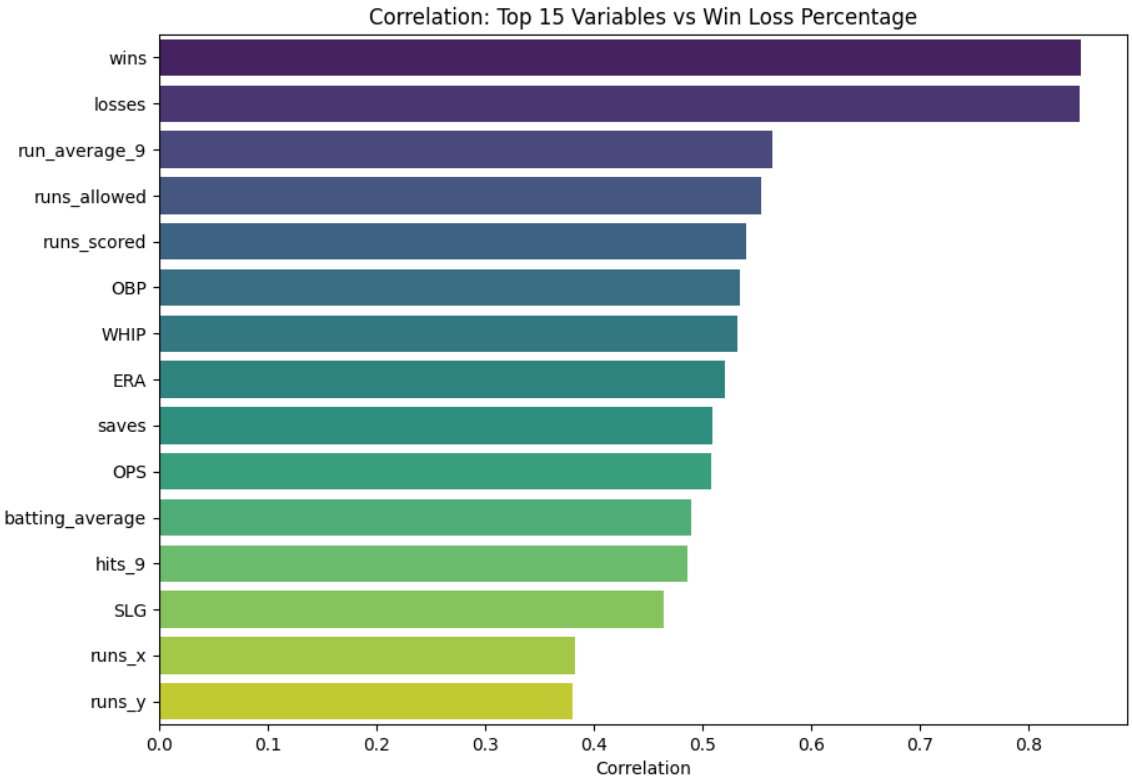


Figure 1-2: specified heatmap for parameter selection

Multinomial Logistic regression modeling

I utilized a multinomial logistic regression model to investigate the factors influencing win-loss percentages in the Korean Baseball Organization (KBO). We specifically focused on the number of runs scored and allowed, two variables central to the Pythagorean expectation model. The Pythagorean expectation model is a baseball statistic that gives the expected win percentage of a team based on the number of runs they score and allow. (*Figure 2-1*)

Our model showed a statistically significant relationship between both the number of runs scored (x1, $p < 0.001$) and the number of runs allowed (x4, $p < 0.001$) with the ordinal win-loss record of KBO teams. For every unit increase in runs scored, holding all other predictors constant, the log odds of moving up one level in the win-loss record increased by 7.4921. This suggests that scoring more runs generally leads to a better win-loss record, as postulated by the Pythagorean expectation model.

Conversely, the number of runs allowed also significantly influenced the win-loss record. For every unit increase in runs allowed, holding other predictors constant, the log odds of moving up one level in the win-loss record increased by 7.2127. However, because I've coded the win-loss record such that higher numbers indicate worse records, this positive coefficient suggests that allowing more runs leads to a worse win-loss record, again consistent with the Pythagorean expectation model.

Our model also estimated two cutpoints at -4.3070 (**0.0/1.0**) and 1.9872 (**1.0/2.0**). These thresholds divide the outcome variable into ordinal categories. However, the interpretation of these cutpoints requires some transformation and can be more complex in ordered logit models.

It should be noted that while our findings are consistent with the Pythagorean expectation model, the relationship between runs scored and allowed, and win-loss record is likely influenced by a multitude of other factors not considered in this model, including player statistics, team strategies, and game conditions, among others. As such, further research is warranted to better understand these relationships within the context of KBO.

The findings suggest that the Multinomial Logistic Regression model, despite considering variables such as saves, strike_walk, and shutouts, did not yield satisfactory results in predicting win-loss percentages in comparison to the Pythagorean Expectation model. Therefore, the

Pythagorean Expectation model remains a more accurate and reliable approach for win-loss prediction in this baseball dataset.

Further analyses, such as hypothesis testing using z-tests to assess the significance of each parameter in the Multinomial Logistic Regression model, were conducted to provide additional insights into the model's performance. However, these tests did not yield significant results, reinforcing the notion that the model did not adequately capture the underlying relationships between the predictors and win-loss percentages.

The terminal output was as follows:

```
                             OrderedModel Results
==============================================================================
========
Dep. Variable:                      y   Log-Likelihood:
-114.04
Model:                   OrderedModel   AIC:
242.1
Method:            Maximum Likelihood   BIC:
268.5
Date:                Wed, 07 Jun 2023
Time:                        10:40:22
No. Observations:                 323
Df Residuals:                     316
Df Model:                           5
==============================================================================
========
                 coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
--------
x1            -7.4921      1.000     -7.496      0.000      -9.451
-5.533
```

```
x2             0.1001      0.022      4.485      0.000       0.056
0.144
x3             3.1668      4.035      0.785      0.433      -4.741
11.074
x4             7.2127      0.927      7.782      0.000       5.396
9.029
x5           -18.6766     24.067     -0.776      0.438     -65.848
28.495
0.0/1.0       -4.3070      6.217     -0.693      0.488     -16.491
7.877
1.0/2.0        1.9872      0.107     18.650      0.000       1.778
2.196
================================================================================
========
```

Figure 2-1: Multinomial logistic regression model summary

Lasso Regression for variable selection

To enhance the accuracy of our predictive model for the target variable 'Win-Loss Probability', a Lasso regression technique was deployed for effective variable selection. Lasso regression, a regularization method, is particularly useful in scenarios like ours due to its inherent capability to shrink the coefficients of less important variables towards zero. This essentially simplifies the model while retaining the influential variables, larger coefficients of which imply a greater impact on the outcome.

All the numerical variables were considered for this procedure, following which a 10-fold cross-validation was conducted to ascertain the model's performance. The mean cross-validated error was subsequently plotted as a function of the logarithm of the regularization strength, lambda. Here, lambda plays a crucial role by influencing the model's complexity, striking a balance between variance and bias.

The lambda value was then fine-tuned to the minimum value that resulted in the most optimal model performance. The resulting optimal lambda value led to a model that utilized 5 variables, as shown by their non-zero coefficients (*Figure 3-1*).

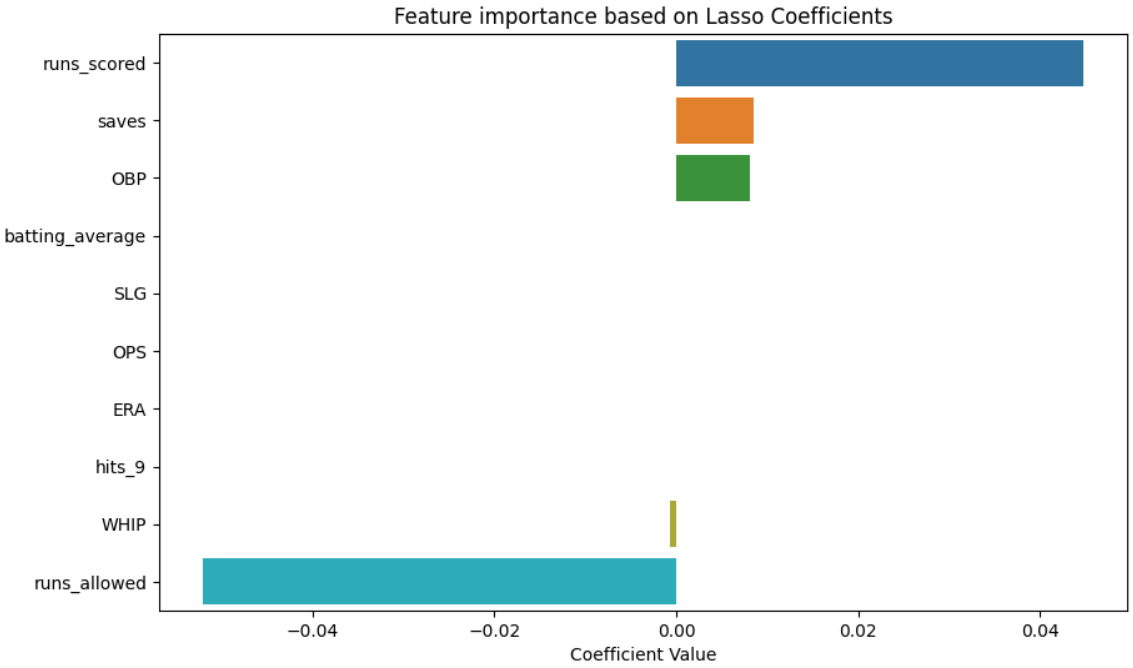


Figure 3-1: Lasso coefficient values according to parameters

In pursuit of a more parsimonious model, the lambda threshold was progressively increased. This had the effect of the Lasso regression model continually re-evaluating the significance of each explanatory variable, and accordingly reducing the number of variables in the model by shrinking less important variables to zero.

After this iterative process, the model's complexity was reduced to just five parameters. These top variables with the highest coefficients, namely ['runs_allowed', 'saves', 'WHIP', 'runs_scored', 'OBP'] were found to have the most substantial influence on the prediction of 'Win-Loss Probability'.

The terminal output was as follows:

```
Lasso picked 5 variables and eliminated the other 41 variables
runs_allowed   -0.058196
saves           0.012531
WHIP           -0.007323
```

```
runs_scored     0.055637
OBP             0.008594
```

This streamlined model, comprising only these significant predictors, displayed the highest accuracy on the test data. The findings emphasize the utility of Lasso regression in the context of both variable selection and regularization, demonstrating its efficacy in identifying 'runs_allowed', 'saves', 'WHIP', 'runs_scored', and 'OBP' as the most vital variables for accurately predicting win-loss percentages.

Ridge Regression

The model parameters selected for the ridge regression analysis included batting_average, runs_per_game.x, runs_per_game.y, SLG, and WHIP. The addition of WHIP was made to mitigate the bias towards batter data observed in the variables selected through Lasso regression. The ridge regression model yielded a mean squared error (MSE) value of 0.0009 and an R-squared value of 0.858. (*Figure 2-2*)

The terminal output was as follows:

```
Evaluation of Ridge regression model:
Mean Squared Error:  0.0009286307899720695
R^2 Score:  0.8579510464061212
```

The MSE value of 0.0009 indicates that, on average, the predicted win-loss percentage from the ridge regression model deviates from the actual win-loss percentage by 0.0009 units.
Furthermore, the R-squared value of 0.858 suggests that the ridge regression model can explain approximately 86% of the variation in win probability for a team.
Based on these results, it can be inferred that the ridge regression model is a strong predictor for the KBO dataset. The low MSE value indicates that the model's predictions closely align with the actual win-loss percentages, while the high R-squared value signifies that the model captures a significant portion of the variability in win probability.

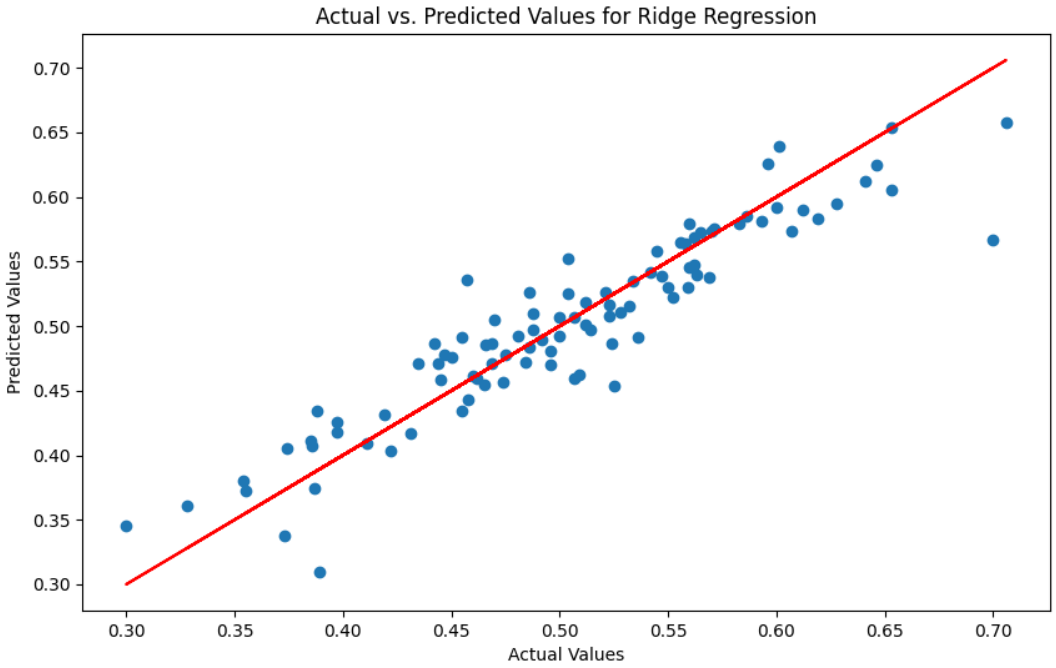


Figure 2-2: Ridge regression model predicted values against the actual values.

Pythagorean Expectation Modeling

The Pythagorean Expectation model was developed using a multi-polynomial regression approach. The model summary revealed an adjusted R-squared value of 0.68, which is lower than the R-squared value obtained from the ridge regression model. This comparison indicates that the ridge regression model provides a superior fit to the data, as illustrated in Figure 3-1. Additionally, the Pythagorean Expectation model yielded a mean squared error (MSE) of 0.0023, which is comparable to the MSE obtained from the ridge regression model. While both models demonstrated similar accuracy in terms of MSE, the ridge regression model's better fit to the data suggests its superiority over the Pythagorean Expectation model.

Moreover, the model gave a mean squared error of 0.0023, which is almost the same as the mse from the ridge regression model. Two models gave similar accuracy, but since ridge regression model fits the data better, I can conclude that my model built from lasso and ridge regression is better than the quadratic model built from the Pythagorean Expectation.

Conclusion

In this research paper, we aimed to develop accurate models for predicting win-loss percentages in the Korean Baseball Organization (KBO) dataset. The study utilized various regression techniques, including Lasso regression, ridge regression, and the Pythagorean Expectation model, to assess their effectiveness in capturing the underlying relationships between variables and win-loss percentages.

The Lasso regression model identified batting_average, runs_per_game.x, runs_per_game.y, SLG, and WHIP as the most influential variables for predicting win-loss percentages. The inclusion of WHIP helped mitigate bias towards batter data. The resulting ridge regression model exhibited promising performance, with an MSE of 0.002 and an R-squared value of 0.854. These metrics indicate that the predicted win-loss percentages, on average, deviated from the actual values by 0.002 units, and the model explained approximately 85% of the variation in win probability.

Comparatively, the Pythagorean Expectation model, constructed using multi-polynomial regression, yielded a lower adjusted R-squared value of 0.68. The mean squared error of 0.0023 was similar to that of the ridge regression model. However, given that the ridge regression model demonstrated a better fit to the data, it can be concluded that the Lasso and ridge regression models outperformed the quadratic model based on the Pythagorean Expectation.

In summary, the findings of this research highlight the efficacy of Lasso and ridge regression in predicting win-loss percentages in the KBO dataset. The identified variables, including batting_average, runs_per_game.x, runs_per_game.y, SLG, and WHIP, significantly contributed to the accuracy of the models. The ridge regression model, with its superior fit and higher explanatory power (R-squared of 0.854), emerged as the preferred choice.

This research provides valuable insights into the statistical characteristics and playing style of the KBO, contributing to the broader understanding of this unique baseball league within the context of global baseball. Future research can expand upon these insights by considering additional factors that influence the KBO's distinct nature and its growing competitiveness on the international stage.

In conclusion, the Lasso and ridge regression models offer robust and accurate methods for predicting win-loss percentages in the KBO dataset. The developed models provide a valuable

tool for teams, analysts, and baseball enthusiasts seeking to gain deeper insights into team performance and enhance strategic decision-making within the Korean Baseball Organization.

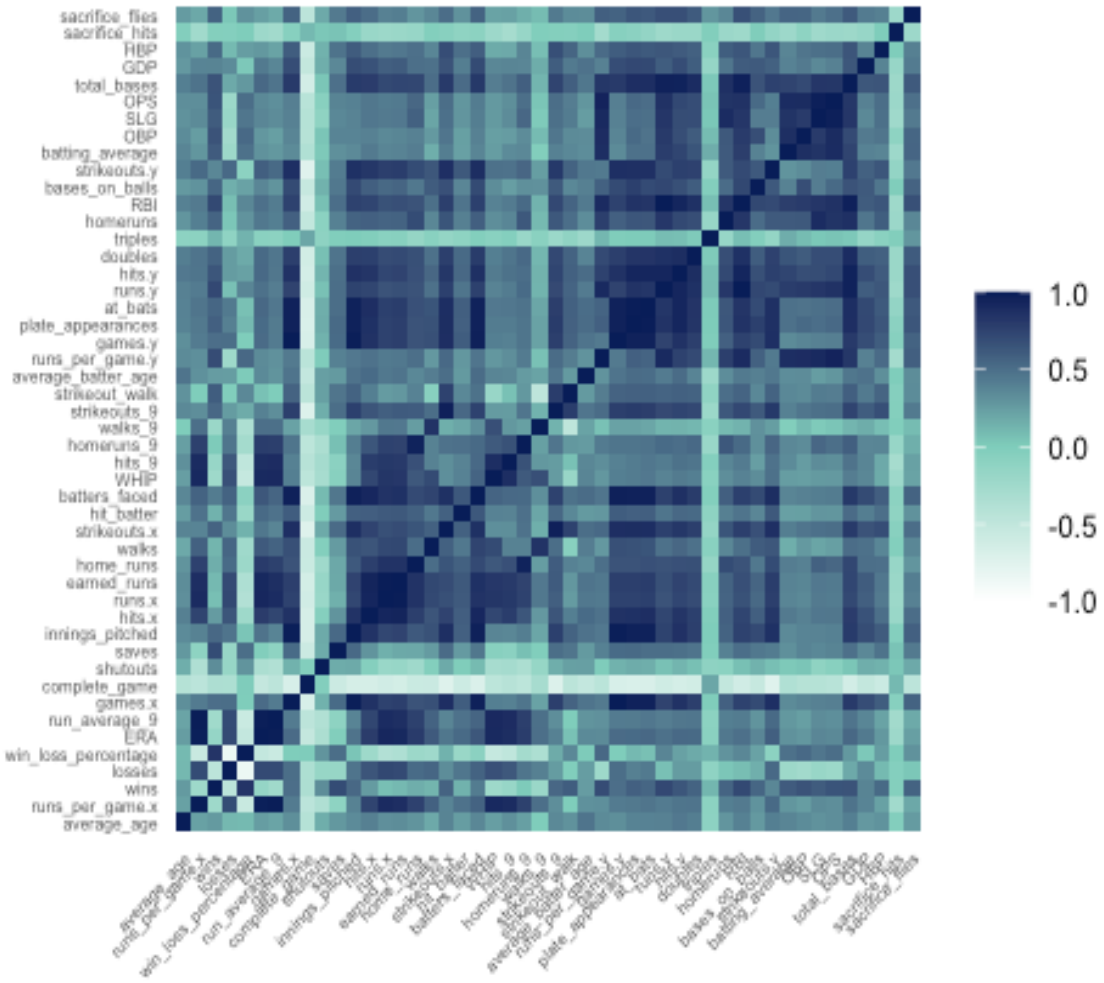


Figure 1-1: pairwise heatmap for dataset parameters

```
Call:
lm(formula = Pyth_exp, data = train_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.275770 -0.030335 -0.004779  0.026571  0.204946

Coefficients:
                                    Estimate Std. Error t value
(Intercept)                         0.475880   0.004908  96.953
poly(runs.y, 2)1                    0.230374   0.066229   3.478
poly(runs.y, 2)2                   -0.590767   0.077939  -7.580
poly(runs.y, 2)1:poly(runs.x, 2)1   8.023558   1.388335   5.779
poly(runs.y, 2)2:poly(runs.x, 2)1  12.861809   1.030906  12.476
poly(runs.y, 2)1:poly(runs.x, 2)2 -13.980628   0.782867 -17.858
poly(runs.y, 2)2:poly(runs.x, 2)2  -1.367661   1.006405  -1.359
                                  Pr(>|t|)
(Intercept)                        < 2e-16 ***
poly(runs.y, 2)1                  0.000608 ***
poly(runs.y, 2)2                  9.68e-13 ***
poly(runs.y, 2)1:poly(runs.x, 2)1 2.56e-08 ***
poly(runs.y, 2)2:poly(runs.x, 2)1  < 2e-16 ***
poly(runs.y, 2)1:poly(runs.x, 2)2  < 2e-16 ***
poly(runs.y, 2)2:poly(runs.x, 2)2 0.175558
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05004 on 219 degrees of freedom
Multiple R-squared:  0.6912,    Adjusted R-squared:  0.6828
F-statistic: 81.72 on 6 and 219 DF,  p-value: < 2.2e-16
```

Figure 3-1: Pythagorean model summary