# Korean Baseball Organization Player Projections

*As of June 2020, the KBO (Korean Baseball Organization) was one of the few sports leagues currently being broadcast in the US. This sparked my interest in the league overall, and my interest in projecting player performance. These projections predict hitter performance in three categories (RBI, HR, and BA) for the full 2020 season based on player performance over the first ~25 games.*

**1. Data**

Data was scraped from the always amazing Baseball Reference: https://www.baseball-reference.com/register/league.cgi?id=17edbc3b

**2. Data Wrangling/Cleaning**

Data Wrangling Notebook

I had to deal with a varitety of data quality issues. Missing values were either dropped or filled with iterative imputer. I coerced some data types. Batting stance was indicated by a special character in the Player Name column, so that had to be extracted.

**3. EDA**

EDA Notebook

I looked for expected or unexpected relationships in the data with a correlation matrix heatmap. Most of the correlations with our response variables were things a baseball fan might suspect, although the exact numerical values of those correlations would be hard to estimate without taking a deeper look at the data.

**4. Modeling and Tuning**

Modeling Notebook

I used some human judgement to reduce multicollinearity in preprocessing (our model doesn't need to count all three of OPS, OBP, and SLG), and then tried a variety of regressors.

*Ridge Regression* ultimately performed the best for each of the three targeted response variables. This made sense as a top contender, because despite my best efforts, multicollinearity was ultimately going to be a major part of the dataset. And it deals with that fairly well.

From there, I tuned each independent model based on each of the response variables, to end up with one separate model for each.

**5. Model Predictions**

Documentation Notebook

The model performed fairly well. Its predicted top five performers outperformed naive predictions (assuming 2019's players would maintain their rankings) by 42%. The top performing player at the end of the season had been predicted at #2.

**6. Future Improvements**

I would love to work with more detailed data if it's available somewhere. At bat or individual game-by-game statistics would make for a more powerful forecaster. A forecast that took into account a player's performance from past seasons would also improve the forecast's accuracy. Forecasting pitcher performance and/or using pitcher data to inform hitter projections would additionally add value to this project. Ultimately, a forecast that predicted things on a smaller scale would be of more value. Game-by-game forecasts, individual at-bats, or even pitch-by-pitch forecasts could be a possibility with more detailed source data.

**7. Project Writeup:** Project Writeup

**8. Credits**

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.