Capstone Project 1: Baseball and Linear Regression
By Scott W. Lew

Milestone Report Capstone

Project Overview

Problem Statement

Our goal is to predict a baseball teams wins for a single season using a combination of batting and pitching statistics. It is assumed that some of these predictors have causation and could identify areas of a team's performance that needs improvement in order to win more games. For instance, if a team's batting average is determined to be a strong predictor of a team's wins, then batting average is an area that could be focused on by a coach or manager in order to improve performance.

Background

Baseball is considered America's pastime, and one of it's oldest games
During a Major League Baseball (MLB) season, a total of 162 games are played in the modern era of the game. Two opposing teams take turns playing offense, batting, and playing defense, pitching and fielding.

Client

The prospective clients are baseball coaches and managers in professional, collegiate, and amateur leagues who are interested in winning more games.

Approach

Batting and pitching statistics from 1876-2018 will be collected from the web and merged into a single table as a Pandas dataframe. Then, regression analysis will be applied to determine the features/variables that are most useful in predicting a team's wins for a season.

Deliverables

The code will be written in the form of Jupyter Notebooks will be displayed and shared on Github. In addition, slides and a written reports will also be available on Github.

Description of Dataset

The data set consists of both batting and pitching statistics for all major league baseball (MLB) teams in the years 1876-2018. The data was scrapped from the website baseball-reference.com using the Beautiful Soup program. A batting statistics table and a pitching statistics table were merged for by joining on team and year columns. Then, tables for each season were concatenated to obtain all the MLB statistics from 1876-2018.

Datasets

Data was collected by web scraping using the python program 'Beautiful Soup'. All pitching and batting statistics were taken from tables available on the website,baseball-reference.com. For this project, statistics were taken for all Major League Baseball (MLB) seasons from 1876 to 2018.

The batting statistics are made up of 28 features while the pitching statistics contain 35 features.

Data Wrangling and Data Cleaning

All batting statistics were downloaded and saved as csv files and then converted into pandas dataframes which were subsequently concatenated. This approach was also used for the pitching statistics data. Then, the dataframes for pitching and batting were merged using an inner join on two columns: Team(Tm) and Year, which produced the entire dataset used for the subsequent analysis.

Exploratory Data Analysis and Initial Findings

Data Features

A total of 29 features were used to predict wins. They are the following:

#Bat: Number of players used in game

BatAge: Batters average age

G_x: Games played by team

PA: Plate appearances

AB: At bats

R_x: Runs scored offensively

H_x: Hits produced by team's batting

2B: Double hits on which the batter reaches second base safely without the contribution of a fielding error.

HR_x: Home runs produced by team's batters.

BB_x: Bases on balls and walks

SO_x: The number of strikeouts made by team's batters

OPS+: A metric consisting of a team's on-base plus slugging percentage and normalizes the number across the entire league.

TB: Total bases made by teams batting, where singles counts as 1 base, doubles as 2 bases, triples as 3 bases and home runs as 4 bases.

#P: The number of pitchers used in games

BB_y: bases given up as walks or balls by team's pitching

BF: Batters faced

BK: Balks. A balk is an illegal act by the pitcher when one or more runners are on base.

CG: Complete games

ER: Earned runs allowed by pitching

G_y: Games pitched

GF: A relief pitcher is credited with a game finished (denoted by GF) if he is the last pitcher to pitch for his team in a game.

H_y: Hits allowed by team's pitching

HR_y: Home runs allowed by team's pitching

IP: Innings pitched

PAge: The average age of the team's pitchers.

R_y: Runs allowed by pitching

SO_y: Strikeouts produced by team's pitching

SV: A save is awarded to the relief pitcher who finishes a game for the winning team, under certain circumstances. It is awarded if the relief pitcher maintains his team's leads and pitches at least 3 innings. It is a metric of how well the team's relief pitchers are performing.

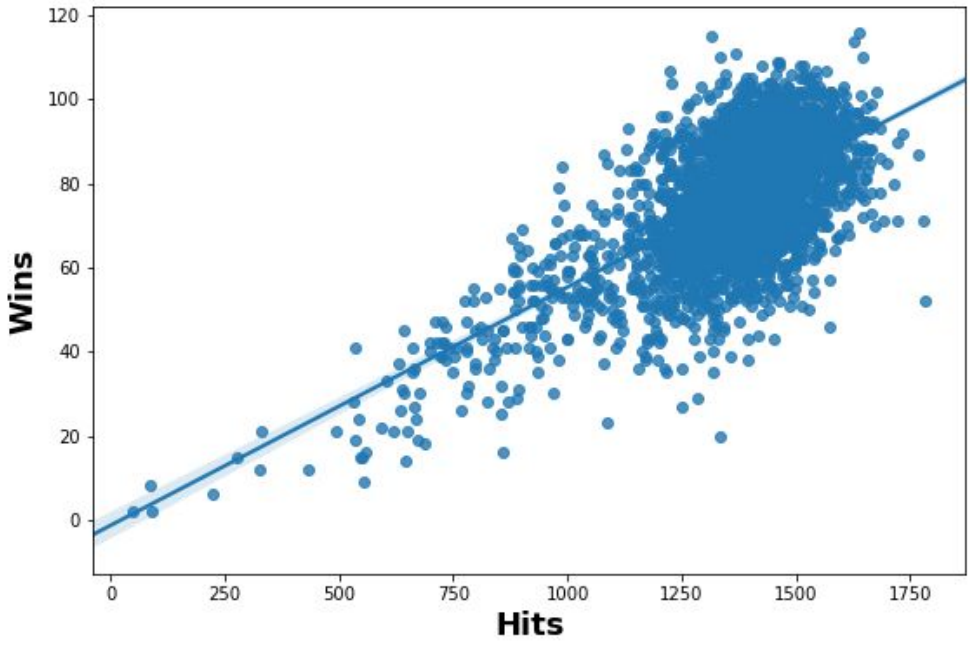
WP: A wild pitch (WP) is made by a pitcher when his pitch is either too high, too short, or too wide of home plate for the catcher to control with ordinary effort, thereby allowing a baserunner, perhaps even the batter-runner on an uncaught third strike, to advance. It is another metric to evaluate the team's pitching ability.

Correlation of Features to Wins

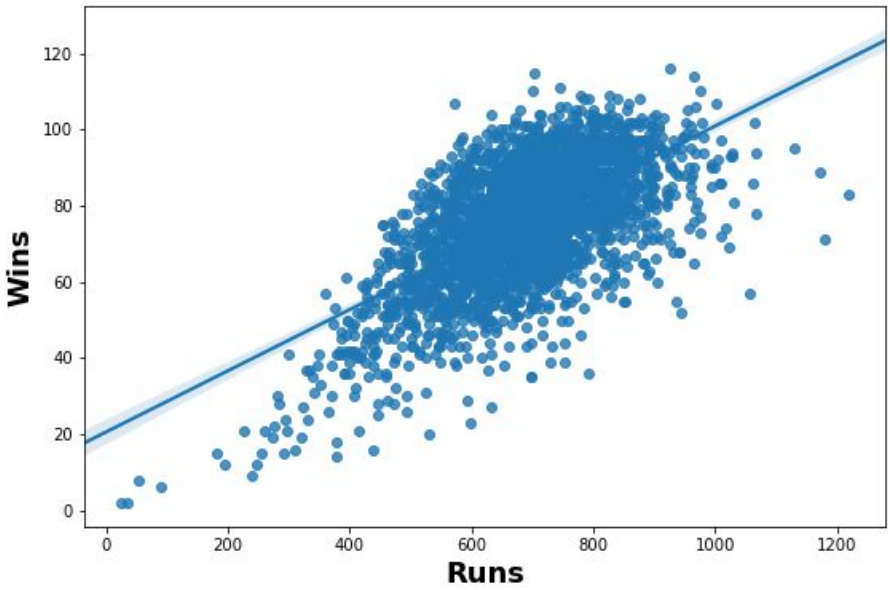
For a linear regression analysis to predict wins, it would be ideal to select those features that have a high correlation with wins.

Scatter plots of several variables showing a positive or negative correlation with the target variable, wins are shown below.

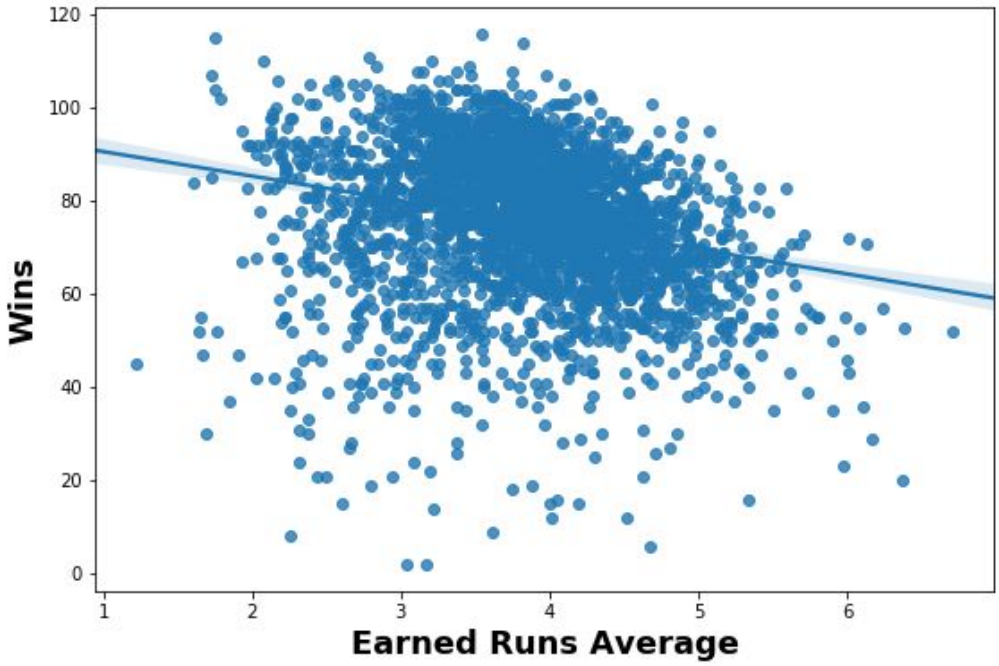
Positive correlation of wins with hits.



As expected, there is a positive correlation of wins with runs produced by hitting.

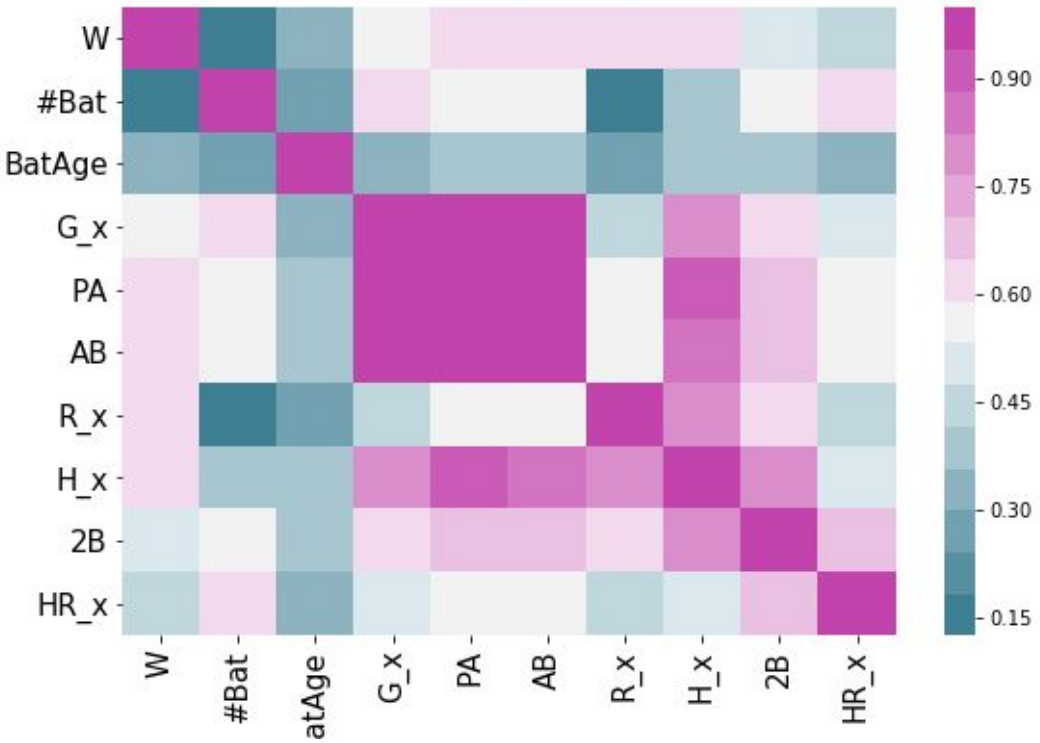


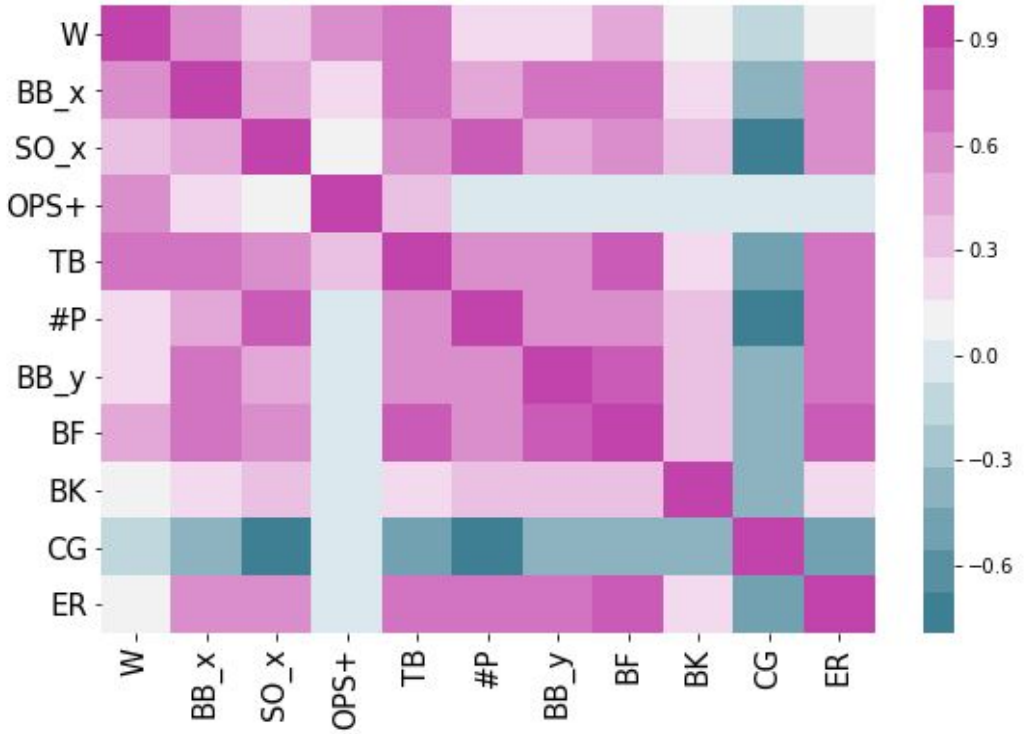
And, there is a negative correlation of Earned Runs Average (ERA) with wins.

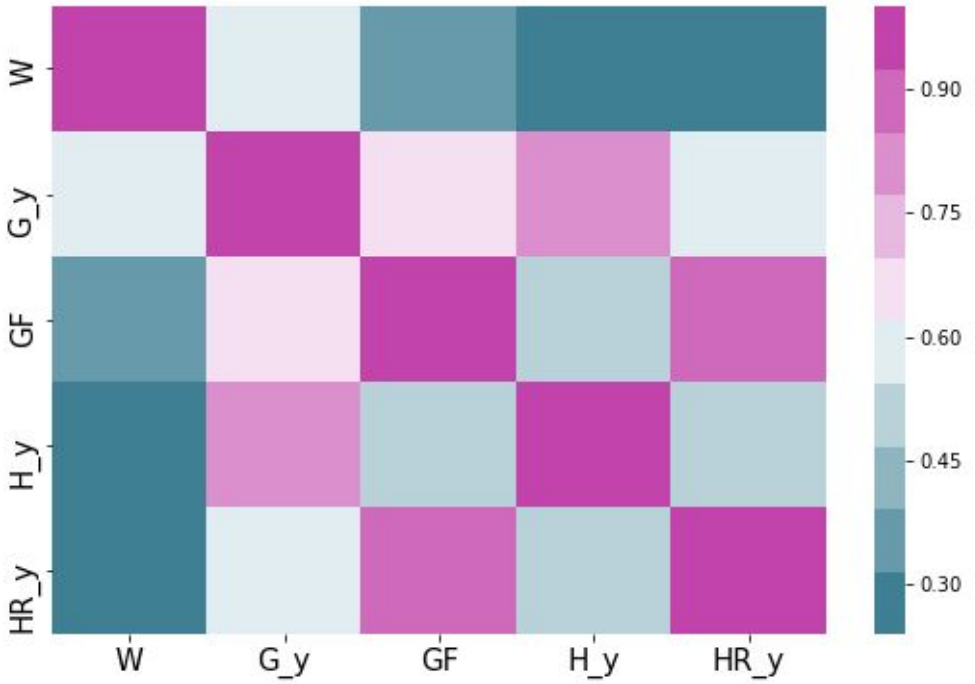


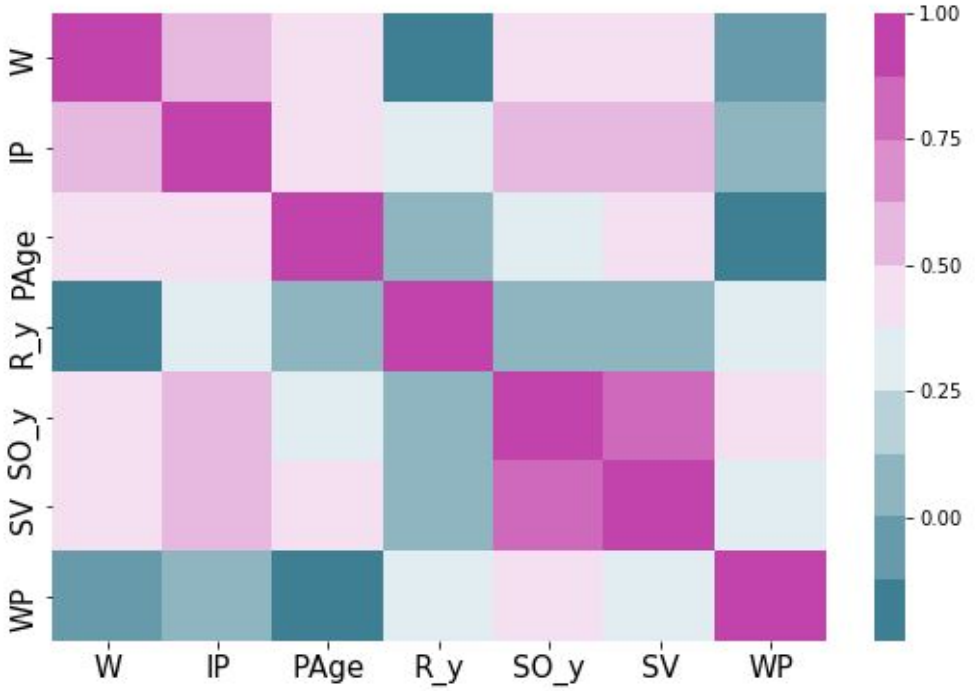
Heatmap of Correlation Matrix of Features with Wins

Another way of visually displaying the correlation of features with one another is by using a heatmap. Features that have a strong positive correlation with wins are shaded more purple, while those features that have a negative correlation have a darker blue color.

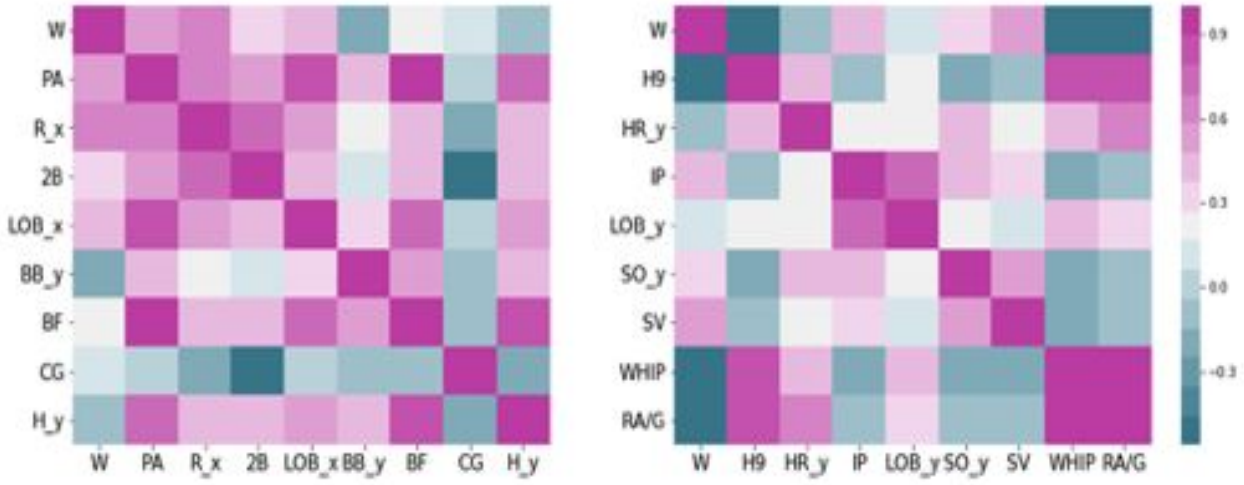



W
BB_x
SO_x
OPS+
TB
#P
BB_y
BF
BK
CG
ER
0.9
0.6
0.3
0.0
−0.3
−0.6


W
G_y
GF
H_y
HR_y
0.90
0.75
0.60
0.45
0.30


W
IP
PAge
R_y
SO_y
SV
WP
W
IP
PAge
R_y
SO_y
SV
WP
1.00
0.75
0.50
0.25
0.00


W
R_x
BA
H_x
SV
ERA
W
R_x
BA
H_x
SV
ERA
1.00
0.75
0.50
0.25
0.00



Independent Variables Positive Correlation With Wins Table 1

Feature	Correlation
TB	**0.649951**
H_x	**0.649828**
RBI	**0.648062**
PA	**0.629993**
R_x	**0.626056**
IP	**0.612138**
AB	**0.599076**
BB_x	**0.586737**
G_y	**0.582832**
G_x	**0.582830**
GS	**0.582657**
OPS	**0.564913**
ERA+	**0.563912**
OPS+	**0.549647**
OBP	**0.533598**
SLG	**0.521056**

As shown in the above table, Total bases(TB), batting hits (H_x), runs scored by batting (R_x), and runs batted in (RBI) all have a positive correlation with a team's wins. Total bases (TB) is the sum of all bases obtained from singles, doubles, triples, and home runs. RBI is defined as a statistic in baseball that measure the total number of runs a hitter generates off of their at-bats with exception to runs scored due to errors by the fielding team.

Independent Variables Negative Correlation With Wins Table 2

Feature	Correlation
RA/G	-0.607445
H9	-0.487337
WHIP	-0.336374
L	-0.332367
ERA	-0.242158
CG	-0.175771
R_y	-0.162247
WP	-0.107353
HBP_y	-0.092135
IBB_y	-0.050684

As shown in the above table, Runs Allowed Per Game(RA/G), and hits allowed per innings pitched (H9) have a negative correlation with a team's wins, which makes perfect sense.

Distribution of wins per season from 1876-2018

The distribution of team wins in a season has a relatively normal distribution that is skewed somewhat.