**Multicollinearity Using Negro League Baseball Data**

The Negro Leagues were made up of American professional baseball teams of African Americans during the late 19th century and into the 1940s. The leagues contained some of the best talent in baseball history and many of which had success in Major League Baseball when the league became integrated in 1947. In 2024, Negro League baseball data was officially incorporated into the record book for Major League Baseball giving credit to some of the game’s best players.

We will be investigating Negro League career batting statistics from 1896 to 1947. The data contains 502 position players (non-pitchers) that competed during this timeframe. Our variables of interest are the number of runs scored, stolen bases, on-base plus slugging plus, homerun percentage, walk-to-strikeout ratio, and runs batted in. A player is credited with a run scored when they successfully make it around the bases and cross Homeplate. A stolen base is when a player runs from the base they are currently at to the next one on a pitch and making it without being tagged out. On-base plus slugging plus is a measure of both how often a player gets and base and accounts for their power (i.e. higher weights given to triples and homeruns than doubles and singles) normalized across the entire league with 100 being the average. Homerun percentage is the number of homeruns hit divided by the number of times the player went up to bat. A walk is when the batter is thrown four balls out of the strike zone when they are up to bat and a strikeout is when the batter gets to three strikes when they are up to bat. Runs batted in is a measure of how many times a batter generated a run by hitting the ball into play.

Our goals will be to predict the number of runs scored by a player using different variables and discovering how strong the relationships between the variables are.

**Data Description**

R – runs scored

RBI – runs batted in

SB – stolen bases

OPS\_plus – on-base percentage plus slugging adjusted for league averages

homerun\_rate – number of homeruns hit divided by plate appearances

BB\_K\_ratio – ratio of walks to strikeouts

1. Before looking at any of the output, which of the variables (besides runs scored) in the data description above do you think will be most correlated with runs scored? Explain.

Sample Answer: I think homerun\_rate would be most correlated with runs scored because when you hit a homerun you are guaranteed to score so I would think a higher rate would more directly lead to more runs scored and vice-versa. None of the other variables guarantee scoring a run.