# Lab 2

Due on 02/16 at 11:59 pm

**Question 1** The 2014 and 2015 Royals surprised a lot of people when they seemingly came out of nowhere with back-to-back world series including a title in 2015. In this problem and in the next problem we will investigate aspects of weirdness surrounding these Royals teams. See this Foolish Baseball video, this Keith Law article, and this article about the failure of projection systems for background. In this problem you will construct a relevant dataset for analysis with the ultimate goal of describing just how unique these Royals were. Do the following:

- Construct a data frame which includes the following variables from the `Teams` data frame in the `Lahman` package: `yearID`, `teamID`, `AB`, `SO`, `H`, `HR`, `R`, `RA`, `W`, and `L`. Only keep seasons dating back to 1990, and remove the 1994, 1995, and 2020 seasons.
- Run the code below to scrape data from baseball reference, and only keep seasons dating back to 1990, and remove the 1994, 1995, and 2020 seasons.

```
bwar_bat = readr::read_csv("https://www.baseball-reference.com/data/war_daily_bat.txt", na = "NULL")
bwar_pit = readr::read_csv("https://www.baseball-reference.com/data/war_daily_pitch.txt", na = "NULL")
```

- Obtain total team defensive WAR `WAR_def`, bullpen WAR, and base running runs `runs_br` for each year and add these quantities to the data frame that you previously constructed from the `Teams` data frame. Call these variables, respectively, `dWAR`, `penWAR`, `BRruns`.
- The 2014-2015 Royals were known for elite base running, an elite bullpen, and elite defense. They were also known for not striking out and not hitting home runs. Add the following scaled variables separately for each season to the data frame that you constructed in the previous step:
  - `scaledSO = scale(SO/AB)`,
  - `scaledBA = scale(H/AB)`,
  - `scaledABpHR = scale(AB/HR)`,
  - `scaledpenWAR = scale(penWAR)`,
  - `scaleddWAR = scale(dWAR)`,
  - `scaledBRruns = scale(BRruns)`
- Compute and add winning percentage `Wpct` to your data frame. Use an equation in your notes and linear regression to compute the optimal $k$ so that `Wpct` is well-explained by `Wpytk` $= R^k/(R^k + RA^k)$. Add `Wpytk` and `residuals_pytk = Wpct - Wpytk` to your data frame.
- Display the rows of this data frame corresponding to the 2014-2015 Royals seasons.

**Question 2** In this problem we will perform analyses that investigate strengths and peculiarities of the 2014-2015 Royals. Do the following:

- Fit and analyze a regression model of `residuals_pytk` on `penWAR`. Determine how many wins one would expect the Royals to obtain above their Pythagorean expectations on the basis of their bullpen.
- Total bullpen WAR is just one aspect of what made the 2014-2015 Royals what they were. We will now use k-means clustering implemented via the kmeans function to determine whether or not teams similar to the 2014-2015 Royals beat their Pythagorean expectations. Do the following with the number of clusters ranging from $k = 30, ..., 50$: 1) run kmeans on a dataset containing the six scaled variables that you previously constructed with $k$ centers; 2) add the cluster assignments to the original dataset; 3) extract the average of `residuals_pytk` for the clusters containing the 2014 or 2015 Royals after

removing the Royals from consideration. When finished, compute the average `residuals_pytk` value for the 2014 and 2015 Royals and then multiply this number by 162. This is the number of expected wins above/below their Pythagorean expectations that similar teams produced. Report this value and compare it with the 2014-2015 Royals.

- Add the `OPSscale` and `WHIPscale` variables that you computed in Question 1 of Lab 1 to the data frame. Run a regression with `Wpct` as the response variable and all eight scaled variables as predictors (you can drop terms if you want to). Does this model over/under estimate the success of the 2014-2015 Royals?

**Question 3** Do the following:

- Select a period of your choice (at least 20 years) and fit the Pythagorean formula model (after finding the optimal exponent) to the run-differential, win-loss data.
- On the basis of your fit in the previous part and the list of managers obtained from Retrosheet, compile a top 10 list of managers who most overperformed their Pythagorean winning percentage and a top 10 list of managers who most underperformed their Pythagorean winning percentage.

**Question 4** The first question on page 21 in Section 1.4.3 of Analyzing Baseball Data with R.