A quantitative study of MLB team performance in April and its relation to season-end results

Electronic version available at
evanmascitti.github.io/april-study/April-study-final-report.pdf

Evan C. Mascitti

Executive summary

Early-season woes are sometimes blamed for a baseball team's poor record.

Losses during the cold spring months accumulate. Indeed, from a mathematical standpoint, these games matter just as much as those later in the year.

Early season performance may also serve to affect a team's psyche. To illustrate the concept, consider an extreme case: the 1988 Orioles, who began 0-21. Enduring a streak which even approaches this level of futility could easily decimate a team's hopes – dooming their season before it has even really begun.

This study compared the records of Major League franchises throughout history. End-of-season results were compared with winning percentages for each month.

Particular attention was devoted to the Philadelphia Phillies. This club is the only professional sports franchise to tally over 10,000 losses, and its generally inferior play has been much maligned by fans and critics alike.

The data suggest that a team's record in April is no better at predicting whether they made the playoffs than their record in other months. In contrast, winning % in the season's later months were better predictors of season-end records.

However, this does not rule out the possibility that a particularly poor April could place an irreparable burden on a team's morale.

For seasons in which the Phillies made the postseason (1976-1978, 1980, 1983, 1993, 2007-2011), their winning percentage often increased over the course of the year. Whether this was due to acquisition of new talent, improved adjustments on the field, or to a more coherent team morale was beyond the scope of this study.

Mathematically, each of the 162 games count equally. However, this study lends some credence to the old adage: "It's not how you start but how you finish."

Materials and methods

Data on 119 MLB seasons were downloaded from the Laman database using the R programming language (R-Core-Team 2021; Friendly et al. 2021).

The data were summarized and plotted using **tidyverse** packages (Wickham et al. 2019), and this report was compiled using R Markdown and GNU `Make` (Allaire et al. 2021; GNU 2020)

MLB-wide study

A broad analysis was first performed over modern MLB history (1903-present). The winning percentage for each team was calculated for each month (April-October) and compared with its final winning percentage (Figure 1).

As one could expect, teams which play well in a given month tend to play well all year long. However, the scatter in Figure 1 appears to diminish later in the year. Perhaps this is a function of good teams improving and adding talent during their playoff push, while poor teams have begun to offload their talent in hopes of rebuilding for the future.

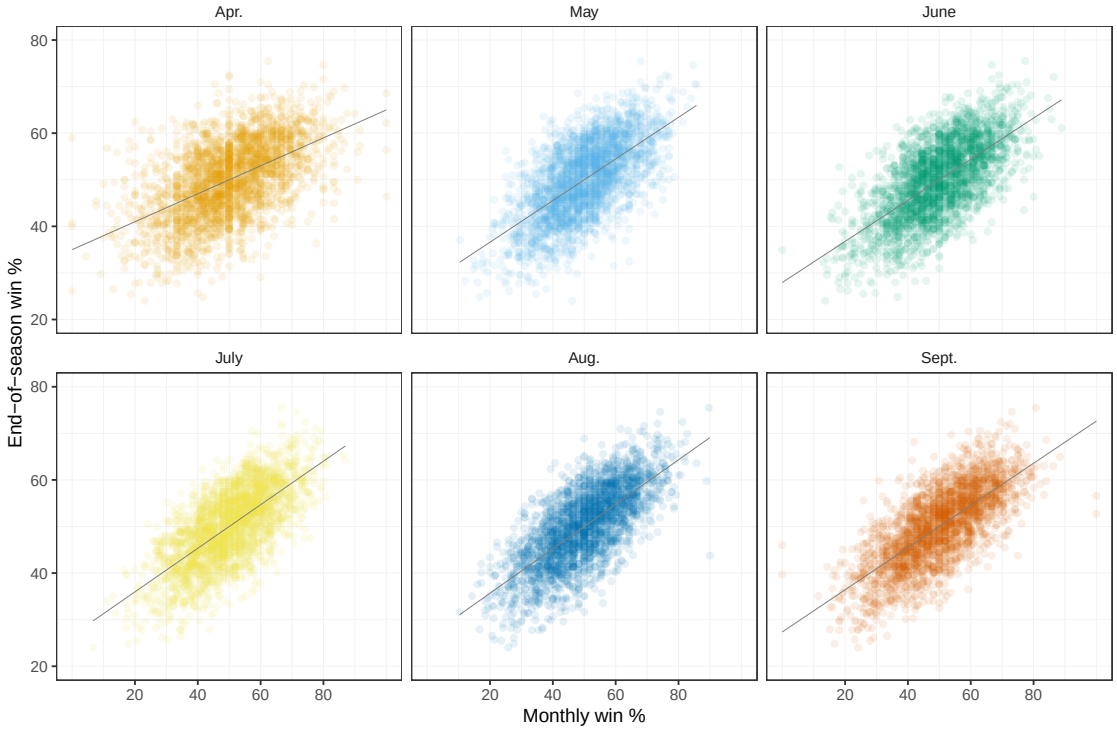

Figure 1: Win percent by month is positively correlated with end-of-season win percent. Scatter appears to decrease later in the season.

Figure 2 shows the strength of the correlations in Figure 1. This suggests that April record has a relatively weak correlation to final record, while a team's record later in the season is a better predictor of their final record. In other words, teams that play their best baseball in August and September tend to also finish with

better records, while advantages gained early in the season tend to be less durable. Teams which play well in April but worse in the later months are unable to “outlast” the teams chasing them.

It must be emphasized that data in Figure 2 bear no conflict with the idea that a good team will play well both at the beginning of the year *and* in later months. It merely shows that April and May seem to have no special importance.

Figure 2: April winnning percentage appears to be less critical than playing well in later months.

Phillies-specific analysis

Figure 3 shows the Phillies daily running winning % for date in the seasons spanning 1970-2019.

These plots tend to be dampened over the course of the season, as the larger number of games played reduces the volatility of the team’ s winning %. Obviously, in the latter months, the effect of winning or losing a single game bears lesser influence on the entire season’s record than it would early in the year, when fewer games have been played.

This makes it difficult to discern any meaningful patterns about the importance of playing well in any given month.

Figure 4 shows an alternative presentation of the same Phillies data.

To reduced the noise in the data and focus on performance *in an individual month*, rather than the cumulative performance up to that date.

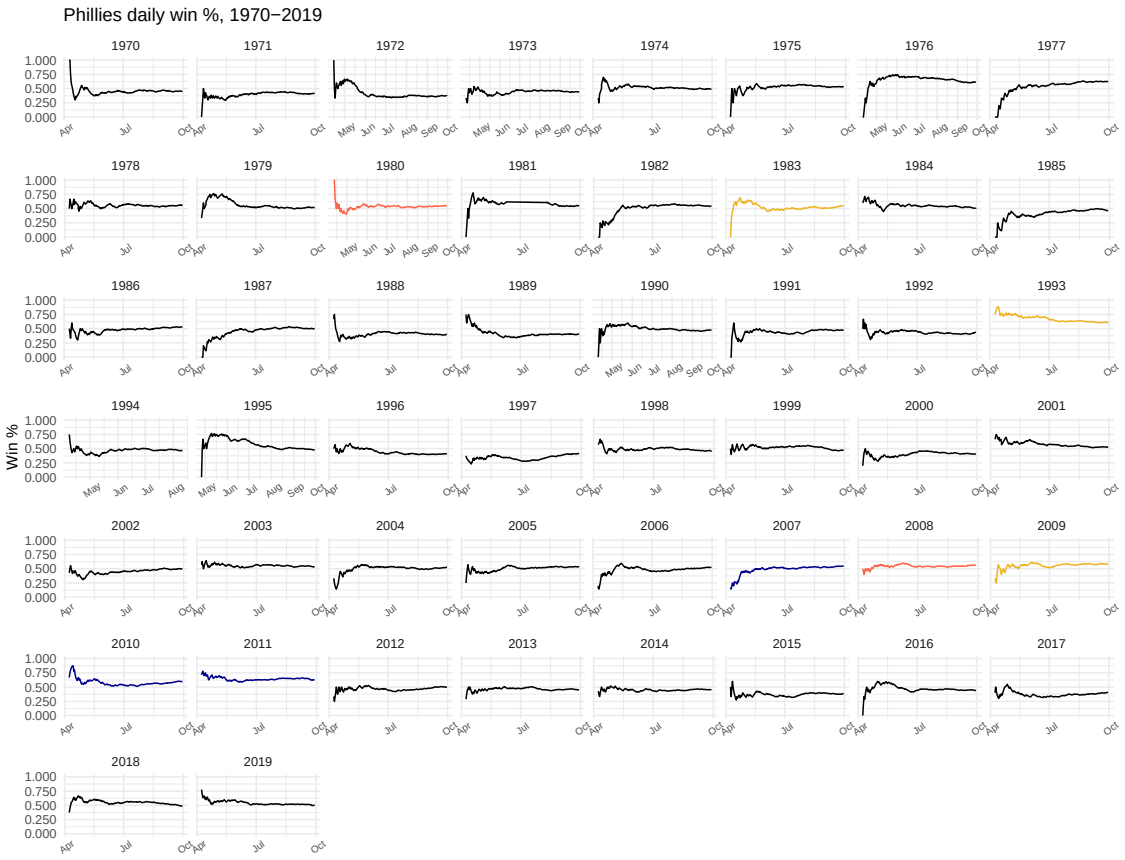

Figure 3: Phillies win percent by day. Red lines indicate World Series championship; yellow lines indicate World Series defeay; blue lines indicate playoff berths.

These are not true time series, because each data point represents the winning percentage for an entire month.

However, they are somewhat more instructive about the influence of a single month on the entire season's result.

For example, in 2008, the Phillies started relatively cold, then experienced an early-summer swoon in May and June. However, the continued to improve throughout the season and caught fire at the perfect time - en route to the World Championship.

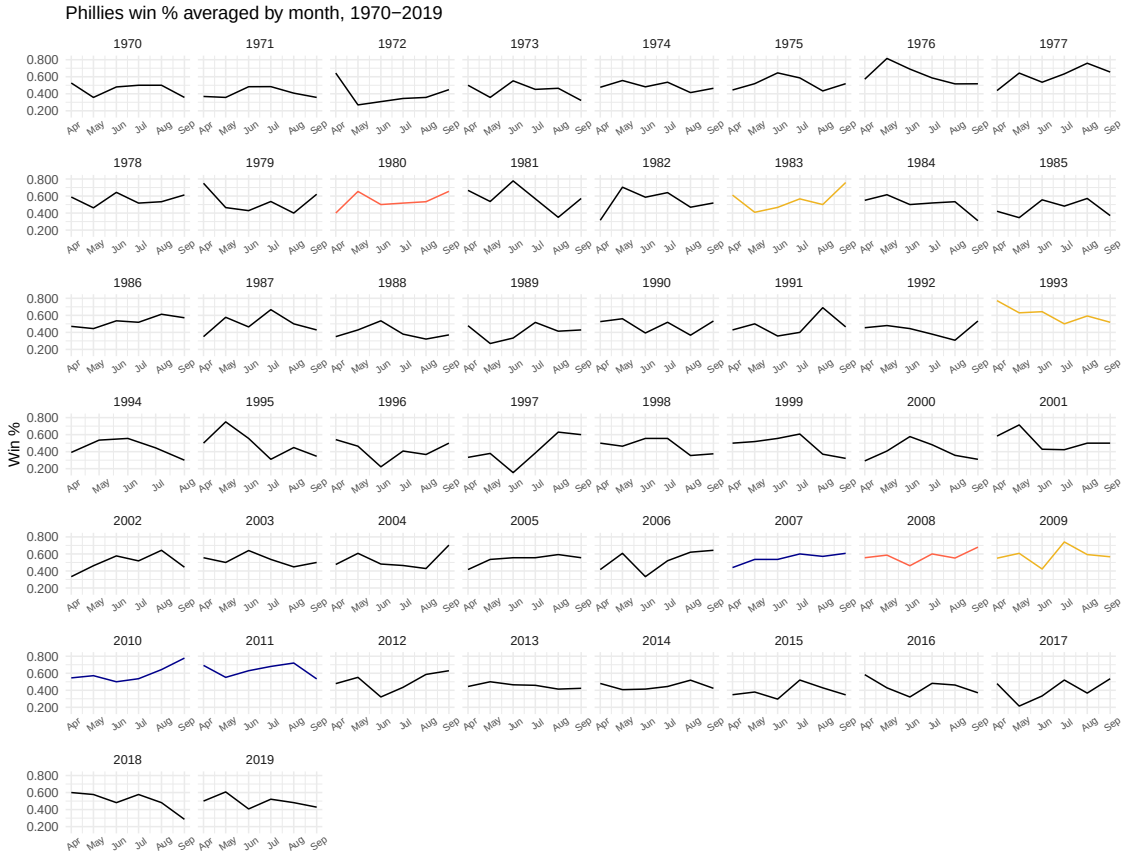

Figure 4: Phillies win percent by month. Red lines indicate World Series championship; yellow lines indicate World Series defeay; blue lines indicate playoff berths.

Conclusions

Making causal inferences about baseball statistics is practically impossible. However, this study identified some interesting trends. Performance in any given month does help predict a team's end-of-season result. Later months appear to have greater explanatory power, but any causal link to team morale is pureley speculative.

In the Phillies most successful seasons, the team often treaded water during the early months and played their best baseball late in the season. This finding should lend hope to Phillies fans, especially in light of the

modern playoff system and its expanded field.

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2021. *Rmarkdown: Dynamic Documents for r*. https://CRAN.R-project.org/package=rmarkdown.

Friendly, Michael, Chris Dalzell, Martin Monkman, and Dennis Murphy. 2021. *Lahman: Sean Lahman Baseball Database*. https://CRAN.R-project.org/package=Lahman.

GNU. 2020. *GNU Make* (version 4.3). The GNU Project.

R-Core-Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.