# Reproducibility checklist

*February 14, 2017*

## Data

### Abstract

Our raw data was purchased from Sports Insights. Each observation is a game in one of the four major North American sports leagues (MLB, NBA, NFL, NHL) between 2005-2016, and contains the game date, the two participating teams, the final score, and the betting lines taken by Sports Insights before the game.

We wrangled these raw data—stripping out duplicate entries, cleaning up team names, etc.—into a single master R `data.frame` called `bigfour.rda`. These data contain the betting lines and therefore cannot be shared. We have placed this file in a private directory denoted `data_raw`.

```
load(file.path(data_raw, "bigfour.rda"))
dim(bigfour)
```

```
[1] 55487    20
```

There are 55,487 observations and 20 variables in `bigfour`.

A derivative public data frame with the same number of observations but many fewer variables is called `bigfour_public` and is available in our GitHub repository.

```
load(file.path("..", "data", "bigfour_public.rda"))
glimpse(bigfour_public)
```

```
Observations: 55,487
Variables: 7
$ gameDate      <dttm> 2013-11-24 13:00:00, 2006-12-24 13:00:00, 2014-...
$ sport         <chr> "nfl", "nfl", "nfl", "nfl", "nfl", "nfl", "nfl",...
$ visitor_team  <chr> "Tampa Bay Buccaneers", "Tampa Bay Buccaneers", ...
$ home_team     <chr> "Detroit Lions", "Cleveland Browns", "Cleveland ...
$ visitor_score <int> 24, 22, 17, 23, 13, 27, 3, 10, 17, 23, 23, 21, 2...
$ home_score    <int> 21, 7, 22, 25, 16, 7, 17, 13, 20, 31, 20, 24, 23...
$ p_home        <dbl> 0.7382366, 0.6447252, 0.7437775, 0.8022976, 0.78...
```

This data frame contains only the game result and the implied probability of the home team winning, based on the betting line.

### Availability

The raw data purchased from Sports Insights (i.e., `bigfour`) is proprietary and cannot be shared publicly. However, we have created a summarized version of these data (i.e., `bigfour_public`) that is publicly shared.

### Description

- Permission to share a version of the Sports Insights data was obtained via email during January, 2017
- Terms of use for Sports Insights data can be found at (https://www.sportsinsights.com/company/terms-of-use/)
- Link to data: (https://github.com/bigfour/competitiveness/data)
- Here's a snapshot of our variables:

| Variable | Description |
|---|---|
| `gameDate` | Day of the game |
| `sport (q)` | League (MLB, NBA, NFL, or NHL) |
| `visitor_team (j)` | Visiting team |
| `home_team (i)` | Home team |
| `visitor_score` | Final score, visiting team |
| `home_score` | Final score, home team |
| `p_home` | $p_{q,i,j}$ |

where $p_{q,i,j}$ is the probability of team home team $i$ defeating away team $j$ in sport $q$.

- Version information: Data purchased on October 6, 2016

## Code

All of our code is hosted in a GitHub repository. The latest commit was made on February 13, 2017. Here are supporting software requirements:

- A recent version of R
- JAGS
- CRAN packages: `tidyverse`, `rjags`, `knitr`, `lubridate`, `stringr`, `xtable`, `XML`, `RCurl`, `zoo`
- GitHub packages: `beanumber/teamcolors`
- `pdflatex`

## Instructions for Use

Please see the README for complete instructions. The first step in reproducing the analysis is to clone the GitHub repository.

> Please note that none of these scripts will actually work without the original proprietary raw data.

- to reproduce the paper, including nearly all figures and tables, type `make` or `make aoas2017.pdf` at the command line
- to reproduce the trace plots, run `Rscript 'R/trace_plots.R'`
- to rerun the MCMC chains, run `make mcmc`
- to reproduce Table 2, run `Rscript 'R/trace_plots.R'` (will only work after the `make mcmc` command)
- to redo the data wrangling, run `make bigfour`
- to recompile the game results, run `make results`

Please do not hesitate to contact us for help reproducing our results.