

## Template Datafile Setup Instructions

For each EPL season from 2014/15 through 2023/24:

1. Copy and paste “Scores & Fixtures” table into a separate Google Sheets for each season on FBRef:

<https://fbref.com/en/comps/9/2014-2015/schedule/2014-2015-Premier-League-Scores-and-Fixtures>  
<https://fbref.com/en/comps/9/2015-2016/schedule/2015-2016-Premier-League-Scores-and-Fixtures>  
<https://fbref.com/en/comps/9/2016-2017/schedule/2016-2017-Premier-League-Scores-and-Fixtures>  
<https://fbref.com/en/comps/9/2017-2018/schedule/2017-2018-Premier-League-Scores-and-Fixtures>  
<https://fbref.com/en/comps/9/2018-2019/schedule/2018-2019-Premier-League-Scores-and-Fixtures>  
<https://fbref.com/en/comps/9/2019-2020/schedule/2019-2020-Premier-League-Scores-and-Fixtures>  
<https://fbref.com/en/comps/9/2020-2021/schedule/2020-2021-Premier-League-Scores-and-Fixtures>  
<https://fbref.com/en/comps/9/2021-2022/schedule/2021-2022-Premier-League-Scores-and-Fixtures>  
<https://fbref.com/en/comps/9/2022-2023/schedule/2022-2023-Premier-League-Scores-and-Fixtures>  
<https://fbref.com/en/comps/9/2023-2024/schedule/2023-2024-Premier-League-Scores-and-Fixtures>

2. Manually delete extra header rows and gray spacing rows
3. Used the following custom AppScript function in Google Sheets to extract URLs from hyperlinked text in “Match Report” column: ([Source](#))

```
function GetURL(input) {  
  var myFormula = SpreadsheetApp.getActiveRange().getFormula();  
  var myAddress = myFormula.replace('=GetURL(', '').replace(')', '');  
  var myRange = SpreadsheetApp.getActiveSheet().getRange(myAddress);  
  return myRange.getRichTextValue().getLinkUrl();  
};
```

4. Used the following command in Google Sheets to split up “Score” into “HomeGoals” and “AwayGoals”

```
=VALUE(LEFT(A2, FIND("-", A2) - 1))  
=VALUE(MID(A2, FIND("-", A2) + 1, LEN(A2) - FIND("-", A2)))
```

5. Enter column names on Google Sheets for variables/statistics to be web scraped from Match Report of each game on FBRef:

<https://fbref.com/en/matches/78c3fc92/Hull-City-Leicester-City-August-13-2016-Premier-League>

6. Export Google Sheets files for each season as an Excel file (.xlsx) and import into R
7. Built and ran web scraper in R (attached as R file)  
(Additional instructions and comments are included in R file)

### Notes:

- Data for the 14/15, 15/16, and 16/17 seasons does not include xG or the number of corners.
- Columns goals.h (17) and goals.a (36) count only goals *scored*, not including own goals
- Attendance column varies across years due to COVID-19 stadium restrictions
- Saves and Save Percentage columns can be N/A in the case that they faced no shot on target