# Read and wrangle Football.co.uk data

## Contents



## Summary

This R-notebook takes as input the raw match data files (one for each season) downloaded from Football.co.uk. It merges them into one dataset, adds a few derived variables, and adds information on which matches were played on artificial turf. Finally, the complete enriched dataset is saved as a R object, ready for analysis.

## Load packages

```
rm(list=ls())
library(data.table)
```

## Read data Eredivisie 2000/2001 - 2017/2018 (up to 2017/12/11)

```
source("code/addTeamIds.R")

# stand 11 december 2017
NL17 <- read.table("data\\N1.csv", header=T, quote="\"", sep=",")
NL16 <- read.table(unz("data\\data16.zip", "N1.csv"), header=T, quote="\"", sep=",")
NL15 <- read.table(unz("data\\data15.zip", "N1.csv"), header=T, quote="\"", sep=",")
NL14 <- read.table(unz("data\\data14.zip", "N1.csv"), header=T, quote="\"", sep=",")
NL13 <- read.table(unz("data\\data13.zip", "N1.csv"), header=T, quote="\"", sep=",")
NL12 <- read.table(unz("data\\data12.zip", "N1.csv"), header=T, quote="\"", sep=",")
NL11 <- read.table(unz("data\\data11.zip", "N1.csv"), header=T, quote="\"", sep=",")
NL10 <- read.table(unz("data\\data10.zip", "N1.csv"), header=T, quote="\"", sep=",")
NL9 <- read.table(unz("data\\data9.zip", "N1.csv"), header=T, quote="\"", sep=",")
NL8 <- read.table(unz("data\\data8.zip", "N1.csv"), header=T, quote="\"", sep=",")
NL7 <- read.table(unz("data\\data7.zip", "N1.csv"), header=T, quote="\"", sep=",", fill = TRUE)
NL6 <- read.table(unz("data\\data6.zip", "N1.csv"), header=T, quote="\"", sep=",")
NL5 <- read.table(unz("data\\data5.zip", "N1.csv"), header=T, quote="\"", sep=",")
NL4 <- read.table(unz("data\\data4.zip", "N1.csv"), header=T, quote="\"", sep=",", fill = TRUE)
NL3 <- read.table(unz("data\\data3.zip", "N1.csv"), header=T, quote="\"", sep=",", fill = TRUE)
```

```
NL2 <- read.table(unz("data\\data2.zip", "N1.csv"), header=T, quote="\"", sep=",", fill = TRUE)
NL1 <- read.table(unz("data\\data1.zip", "N1.csv"), header=T, quote="\"", sep=",", fill = TRUE)
NL0 <- read.table(unz("data\\data0.zip", "N1.csv"), header=T, quote="\"", sep=",", fill = TRUE)


select_cols <- c("Div"   ,    "Date"   ,    "HomeTeam"  , "AwayTeam"  , "FTHG"  ,
 "FTAG"   ,    "FTR"     ,   "WHH", "WHD", "WHA")

NL <- rbind(NL17[, select_cols],
            NL16[, select_cols],
            NL15[, select_cols],
            NL14[, select_cols],
            NL13[, select_cols],
            NL12[, select_cols],
            NL11[, select_cols],
            NL10[, select_cols],
            NL9[, select_cols],
            NL8[, select_cols],
            NL7[, select_cols],
            NL6[, select_cols],
            NL5[, select_cols],
            NL4[, select_cols],
            NL3[, select_cols],
            NL2[, select_cols],
            NL1[, select_cols],
            NL0[, select_cols])

NL <- data.table(NL)
```

## Variable descriptives

Key to results data:

- Div = League Division
- Date = Match Date (dd/mm/yy)
- HomeTeam = Home Team
- AwayTeam = Away Team
- FTHG and HG = Full Time Home Team Goals
- FTAG and AG = Full Time Away Team Goals
- FTR and Res = Full Time Result (H=Home Win, D=Draw, A=Away Win)
- WHH = William Hill home win odds
- WHD = William Hill draw odds
- WHA = William Hill away win odds

## Enrich data DPL

```
NL <- NL[, WinningTeam := "Draw"]
NL <- NL[FTR == "H", WinningTeam := HomeTeam]
NL <- NL[FTR == "A", WinningTeam := AwayTeam]
# add year
```

```
NL <- NL[, Year := as.integer(paste("20", substr(Date, 7,8), sep = ''))]

# drop empty records
NL <- NL[Year != 20,]

# trim whitespace
NL <- NL[, HomeTeam := trimws(HomeTeam, "r")]

# fix team names
NL <- NL[HomeTeam == "Sparta Rotterdam", HomeTeam := "Sparta"]
NL <- NL[AwayTeam == "Sparta Rotterdam", AwayTeam := "Sparta"]
NL <- NL[HomeTeam == "Roda", HomeTeam := "Roda JC"]
NL <- NL[AwayTeam == "Roda", AwayTeam := "Roda JC"]

# add goal difference for each match
NL <- NL[, goal_difference := FTHG - FTAG]


NL <- addTeamIds(NL)
# Eyeball teams in time
#res <- NL[, .N, .(HomeTeam, Year)]
#res <- NL[, .N, .(AwayTeam, Year)]
```

## Add Artificial turf data

Teams with Artificial turf in the DPL: * Heracles (2003) * Excelsior (from 2010), * PEC Zwolle (from 2012) * SC Cambuur (2013) * ADO Den Haag (okt 2013, first four games on natural turf) * Roda JC (2014) * FC Dordrecht (2014) * Sparta (from 2014/2015) * VVV-Venlo (from 2013 in jupiler, 2017)

*Tabel 1 Kunstgrasclubs in de Eredivisie en relatieve ECI score*, 2013/2014 – 2015/2016*

| Kunstgrasclubs | Relatieve ECI-score | | |
|---|---|---|---|
| | 2013/2014 | 2014/2015 | 2015/2016 |
| ADO Den Haag** | .9340 | .9221 | .9428 |
| Excelsior | --*** | .7387 | .7330 |
| FC Dordrecht | -- | .6371 | -- |
| Heracles Almelo | .9022 | .8972 | .9829 |
| PEC Zwolle | .8988 | .9876 | .9677 |
| Roda JC | -- | -- | .8051 |
| SC Cambuur | .7759 | .8438 | .6939 |

*De relatieve ECI score van een team is gemeten door de ECI score van een club te delen door de gemiddelde ECI score van alle Eredivisieteams. Een score kleiner dan 1 geeft dus aan dat een team een lagere ECI score heeft dan het Eredivisiegemiddelde, en een score hoger dan 1 geeft aan dat een team beter is dan het Eredivisiegemiddelde.
**ADO Den Haag begon in het seizoen 2013/2014 op natuurgras, maar stapte in oktober 2013 over op kunstgras. ADO speelde dat jaar tegen PSV, Roda JC, NEC, en Vitesse op natuurgras. De andere wedstrijden zijn op kunstgras gespeeld
***Teams waarbij de relatieve ECI score niet vermeld staat speelden dat seizoen niet in de Eredivisie

Figure 1:

```r
NL <- NL[, Date := as.Date(Date, "%d/%m/%y")]

NL <- NL[, art_turf := 0]
NL <- NL[, art_turf_away := 0]
# * Heracles (2003)
NL <- NL[HomeTeam == "Heracles" & Date > as.Date("2003-06-01"), art_turf := 1]
NL <- NL[AwayTeam == "Heracles" & Date > as.Date("2003-06-01"), art_turf_away := 1]
# * Excelsior (from 2010),
NL <- NL[HomeTeam == "Excelsior" & Date > as.Date("2010-06-01"), art_turf := 1]
NL <- NL[AwayTeam == "Excelsior" & Date > as.Date("2010-06-01"), art_turf_away := 1]
# * PEC Zwolle (from 2012)
NL <- NL[HomeTeam == "Zwolle" & Date > as.Date("2012-06-01"), art_turf := 1]
NL <- NL[AwayTeam == "Zwolle" & Date > as.Date("2012-06-01"), art_turf_away := 1]
# * SC Cambuur (2013)
NL <- NL[HomeTeam == "Cambuur" & Date > as.Date("2013-06-01"), art_turf := 1]
NL <- NL[AwayTeam == "Cambuur" & Date > as.Date("2013-06-01"), art_turf_away := 1]
# * ADO Den Haag (okt 2013, first four games on natural turf)
NL <- NL[HomeTeam == "Den Haag" & Date > as.Date("2013-10-01"), art_turf := 1]
NL <- NL[AwayTeam == "Den Haag" & Date > as.Date("2013-10-01"), art_turf_away := 1]
# * Roda JC (2014)
NL <- NL[HomeTeam == "Roda JC" & Date > as.Date("2014-6-01"), art_turf := 1]
NL <- NL[AwayTeam == "Roda JC" & Date > as.Date("2014-6-01"), art_turf_away := 1]
# * FC Dordrecht (2014)
NL <- NL[HomeTeam == "Dordrecht" & Date > as.Date("2014-6-01"), art_turf := 1]
NL <- NL[AwayTeam == "Dordrecht" & Date > as.Date("2014-6-01"), art_turf_away := 1]
# * Sparta (from 2014/2015)
NL <- NL[HomeTeam == "Sparta" & Date > as.Date("2014-6-01"), art_turf := 1]
NL <- NL[AwayTeam == "Sparta" & Date > as.Date("2014-6-01"), art_turf_away := 1]
# * VVV-Venlo (from 2013 in jupiler, 2017)
NL <- NL[HomeTeam == "VVV Venlo" & Date > as.Date("2013-6-01"), art_turf := 1]
NL <- NL[AwayTeam == "VVV Venlo" & Date > as.Date("2013-6-01"), art_turf_away := 1]

NL <- NL[, art_turf_advantage := 0]
NL <- NL[art_turf == 1 & art_turf_away == 0, art_turf_advantage := 1]

# Vars van Ours & Teijl
NL <- NL[, nat2 := 0]
NL <- NL[art_turf == 0 & art_turf_away == 0, nat2 := 1]

NL <- NL[, natkun := 0]
NL <- NL[art_turf == 0 & art_turf_away == 1, natkun := 1]

NL <- NL[, kunnat := 0]
NL <- NL[art_turf == 1 & art_turf_away == 0, kunnat := 1]

NL <- NL[, kun2 := 0]
NL <- NL[art_turf == 1 & art_turf_away == 1, kun2 := 1]

NL <- NL[, match_type := "Nat-Nat"]
NL <- NL[natkun == 1, match_type := "Nat-Kun"]
NL <- NL[kunnat == 1, match_type := "Kun-Nat"]
NL <- NL[kun2 == 1, match_type := "Kun-Kun"]
```

```r
saveRDS(NL, "data\\NL Eredivisie 2000-2018.rds")
```