clean_data

Xavier Loffree

2024-12-06

```
library(maps)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ---------------------------------------- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::map()    masks maps::map()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(mclust)
```

```
## Package 'mclust' version 6.1.1
## Type 'citation("mclust")' for citing this R package in publications.
##
## Attaching package: 'mclust'
##
## The following object is masked from 'package:purrr':
##
##     map
##
## The following object is masked from 'package:maps':
##
##     map
```

```
library(xgboost)
```

```
##
## Attaching package: 'xgboost'
##
## The following object is masked from 'package:dplyr':
##
##     slice
```

```
library(nnet)
library(glmnet)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
```

```
##
## The following objects are masked from ’package:tidyr’:
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-8
```

```r
library(data.table)
```

```
##
## Attaching package: ’data.table’
##
## The following objects are masked from ’package:lubridate’:
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from ’package:dplyr’:
##
##     between, first, last
##
## The following object is masked from ’package:purrr’:
##
##     transpose
```

```r
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: ’caret’
##
## The following object is masked from ’package:purrr’:
##
##     lift
```

```r
#setwd("C:/Users/xavie/OneDrive/Documents/bios611/project/")
df <- read.csv("Fifa_world_cup_matches.csv", header=TRUE)
```

Clean data and add variables

```r
#Make sure corresponding var names for teams 1 and 2 differ only by the team number
#In other words, some of the column names have typos that we need to fix
colnames(df)[which(colnames(df)=="attempts.inside.the.penalty.area..team2")] <-
  "attempts.inside.the.penalty.area.team2"
colnames(df)[which(colnames(df)=="completed.line.breaksteam1")] <-
  "completed.line.breaks.team1"
colnames(df)[which(colnames(df)=="completed.defensive.line.breaksteam1")] <-
  "completed.defensive.line.breaks.team1"




#Add total goals column
df$total.goals <- df$number.of.goals.team1 + df$number.of.goals.team2

#Add total attempts column
```

```r
df$total.attempts <- df$total.attempts.team1 + df$total.attempts.team2

#Add total attempted line breaks column
df$total.attempted.defensive.line.breaks <-
  df$attempted.defensive.line.breaks.team1 +
  df$attempted.defensive.line.breaks.team2

#Add indicator variable for if game is an elimination game
df$elimination <- as.factor(c(rep(0,48), rep(1,16)))

#Convert percentages to numerical vars
pct_cols <- c("possession.team1", "possession.team2", "possession.in.contest")
df[,pct_cols] <- lapply(df[,pct_cols], function(x) as.numeric(gsub("%", "", x)))

#Add match outcome variable, 1=team1 win, 2=team2 win, 0=tie
df$outcome <- as.factor(ifelse(df$number.of.goals.team1>df$number.of.goals.team2,1,
                     ifelse(df$number.of.goals.team1<df$number.of.goals.team2,2,0)))


write.csv(
  df, "match_data_clean.csv",
          row.names = FALSE)
```