Pràctica 2: Creació de la visualització i lliurament del projecte

VisualiTzació de dades, Universitat Oberta de Catalunya

Juan Luis Espinoza López

13 June 2022

Contents

Anàlisi exploratòria

Primer de tot instal · lem i carreguem les llibreries ggplot2 i dplry.

```
# https://cran.r-project.org/web/packages/ggplot2/index.html
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
# https://cran.r-project.org/web/packages/dplyr/index.html
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
# https://cran.r-project.org/web/packages/dplyr/index.html
if (!require('stringr')) install.packages('stringr'); library('stringr')
```

El primer pas per realitzar un anàlisi exploratòria es carreggar els fitxer de dades que anem a utilitzar

```
path_games = '/Users/jespinlo10/Documents/Master/2on Semestre/Visualizacion/PRAC1/archive/games.csv'
games <- read.csv(path_games,stringsAsFactors = FALSE)
rows=dim(games)[1]
```

```
structure = str(games)
```

```
## 'data.frame':    12680 obs. of  34 variables:
##  $ gameID           : int  81 82 83 84 85 86 87 88 89 90 ...
##  $ leagueID         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ season           : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ date             : chr  "2015-08-08 15:45:00" "2015-08-08 18:00:00" "2015-08-08 18:00:00" "2015-08
##  $ homeTeamID       : int  89 73 72 75 79 80 86 83 85 76 ...
##  $ awayTeamID       : int  82 71 90 77 78 84 74 81 87 88 ...
##  $ homeGoals        : int  1 0 2 4 1 2 2 0 0 0 ...
##  $ awayGoals        : int  0 1 2 2 3 2 2 2 1 3 ...
##  $ homeProbability  : num  0.284 0.357 0.299 0.642 0.146 ...
##  $ drawProbability  : num  0.4 0.35 0.434 0.206 0.216 ...
##  $ awayProbability  : num  0.316 0.293 0.268 0.152 0.638 ...
##  $ homeGoalsHalfTime: int  1 0 0 3 0 2 1 0 0 0 ...
##  $ awayGoalsHalfTime: int  0 0 1 0 1 1 1 1 0 2 ...
##  $ B365H            : num  1.65 2 1.7 1.95 2.55 1.36 2.88 1.29 3.4 5.75 ...
##  $ B365D            : num  4 3.6 3.9 3.5 3.3 5 3.3 6 3.4 4 ...
##  $ B365A            : num  6 4 5.5 4.33 3 11 2.7 12 2.3 1.67 ...
##  $ BWH              : num  1.65 2 1.7 2 2.6 1.4 2.8 1.28 3.2 4.75 ...
##  $ BWD              : num  4 3.3 3.5 3.3 3.2 4.75 3.1 5.75 3.4 4 ...
##  $ BWA              : num  5.5 3.7 5 3.75 2.7 9 2.75 10.5 2.3 1.65 ...
##  $ IWH              : num  1.65 2.1 1.7 2 2.4 1.33 2.65 1.33 2.9 5.1 ...
##  $ IWD              : num  3.6 3.3 3.6 3.3 3.2 4.8 3.3 4.8 3.3 3.6 ...
##  $ IWA              : num  5.1 3.3 4.7 3.6 2.85 8.3 2.5 8.3 2.3 1.65 ...
##  $ PSH              : num  1.65 1.95 1.7 1.99 2.52 1.39 2.88 1.31 3.48 5.75 ...
##  $ PSD              : num  4.09 3.65 3.95 3.48 3.35 4.92 3.33 5.75 3.46 3.98 ...
##  $ PSA              : num  5.9 4.27 5.62 4.34 3.08 ...
##  $ WHH              : num  1.62 1.91 1.73 2 2.6 1.4 2.7 1.3 3.3 5.5 ...
##  $ WHD              : num  3.6 3.5 3.5 3.1 3.1 4 3.1 5 3.1 3.5 ...
##  $ WHA              : num  6 4 5 2.7 2.88 10 2.7 11 2.3 1.7 ...
##  $ VCH              : num  1.67 2 1.73 2 2.6 1.4 2.88 1.3 3.4 5.5 ...
##  $ VCD              : num  4 3.5 3.9 3.4 3.25 5 3.25 5.75 3.4 4 ...
##  $ VCA              : num  5.75 4.2 5.4 4.33 3 9.5 2.7 12 2.3 1.7 ...
##  $ PSCH             : num  1.64 1.82 1.75 1.79 2.46 1.37 3.09 1.24 3.89 6.46 ...
##  $ PSCD             : num  4.07 3.88 3.76 3.74 3.39 5.04 3.28 6.75 3.51 4.08 ...
##  $ PSCA             : num  6.04 4.7 5.44 5.1 3.14 ...
```

```
path_leagues = '/Users/jespinlo10/Documents/Master/2on Semestre/Visualizacion/PRAC1/archive/leagues.csv
leagues <- read.csv(path_leagues,stringsAsFactors = FALSE)
rows=dim(leagues)[1]
```

```
structure = str(leagues)
```

```
## 'data.frame':    5 obs. of  3 variables:
##  $ leagueID         : int  1 2 3 4 5
##  $ name             : chr  "Premier League" "Serie A" "Bundesliga" "La Liga" ...
##  $ understatNotation: chr  "EPL" "Serie_A" "Bundesliga" "La_liga" ...
```

```
path_players = '/Users/jespinlo10/Documents/Master/2on Semestre/Visualizacion/PRAC1/archive/players.csv
players <- read.csv(path_players,stringsAsFactors = FALSE)
rows=dim(players)[1]
```

```
structure = str(players)
```

```
## 'data.frame':    7659 obs. of  2 variables:
##  $ playerID: int  560 557 548 628 1006 551 654 554 555 631 ...
##  $ name    : chr  "Sergio Romero" "Matteo Darmian" "Daley Blind" "Chris Smalling" ...
```

```
path_teams= '/Users/jespinlo10/Documents/Master/2on Semestre/Visualizacion/PRAC1/archive/teams.csv'
teams <- read.csv(path_teams, stringsAsFactors = FALSE)
rows=dim(teams)[1]
```

```
structure = str(teams)
```

```
## 'data.frame':    146 obs. of  2 variables:
##  $ teamID: int  71 72 74 75 76 77 78 80 81 82 ...
##  $ name  : chr  "Aston Villa" "Everton" "Southampton" "Leicester" ...
```

```
path_shots= '/Users/jespinlo10/Documents/Master/2on Semestre/Visualizacion/PRAC1/archive/shots.csv'
shots <- read.csv(path_shots, stringsAsFactors = FALSE)
rows=dim(shots)[1]
```

```
structure = str(shots)
```

```
## 'data.frame':    324543 obs. of  11 variables:
##  $ gameID    : int  81 81 81 81 81 81 81 81 81 81 ...
##  $ shooterID : int  554 555 554 554 555 555 631 629 629 646 ...
##  $ assisterID: int  NA 631 629 NA 654 629 NA 557 NA 647 ...
##  $ minute    : int  27 27 35 35 40 49 64 72 76 4 ...
##  $ situation : chr  "DirectFreekick" "SetPiece" "OpenPlay" "OpenPlay" ...
##  $ lastAction: chr  "Standard" "Pass" "Pass" "Tackle" ...
##  $ shotType  : chr  "LeftFoot" "RightFoot" "LeftFoot" "LeftFoot" ...
##  $ shotResult: chr  "BlockedShot" "BlockedShot" "BlockedShot" "MissedShots" ...
##  $ xGoal     : num  0.1043 0.0643 0.0572 0.0921 0.0357 ...
##  $ positionX : num  0.794 0.86 0.843 0.848 0.812 ...
##  $ positionY : num  0.421 0.627 0.333 0.533 0.707 ...
```

```
path_appearances= '/Users/jespinlo10/Documents/Master/2on Semestre/Visualizacion/PRAC1/archive/appearanc
appearances <- read.csv(path_appearances, stringsAsFactors = FALSE)
rows=dim(appearances)[1]
```

```
structure = str(appearances)
```

```
## 'data.frame':    356513 obs. of  19 variables:
##  $ gameID       : int  81 81 81 81 81 81 81 81 81 81 ...
##  $ playerID     : int  560 557 548 628 1006 551 654 554 555 631 ...
##  $ goals        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ownGoals     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ shots        : int  0 0 0 0 0 0 0 3 3 1 ...
##  $ xGoals       : num  0 0 0 0 0 ...
##  $ xGoalsChain  : num  0 0.1065 0.1277 0.1065 0.0212 ...
##  $ xGoalsBuildup: num  0 0.1065 0.1277 0.1065 0.0212 ...
##  $ assists      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ keyPasses    : int  0 1 0 0 0 0 1 0 0 1 ...
##  $ xAssists     : num  0 0.107 0 0 0 ...
##  $ position     : chr  "GK" "DR" "DC" "DC" ...
##  $ positionOrder: int  1 2 3 3 4 7 7 11 12 13 ...
##  $ yellowCard   : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ redCard      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ time         : int  90 82 90 90 90 90 61 90 69 90 ...
##  $ substituteIn : int  0 222605 0 0 0 0 222606 0 222607 0 ...
##  $ substituteOut: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ leagueID     : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
path_teamstats= '/Users/jespinlo10/Documents/Master/2on Semestre/Visualizacion/PRAC1/archive/teamstats.
teamstats <- read.csv(path_teamstats, stringsAsFactors = FALSE)
rows=dim(teamstats)[1]
```

```
structure = str(teamstats)
```

```
## 'data.frame':    25360 obs. of  16 variables:
##  $ gameID       : int  81 81 82 82 83 83 84 84 85 85 ...
##  $ teamID       : int  89 82 73 71 72 90 75 77 79 78 ...
##  $ season       : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ date         : chr  "2015-08-08 15:45:00" "2015-08-08 15:45:00" "2015-08-08 18:00:00" "2015-08-08
##  $ location     : chr  "h" "a" "h" "a" ...
##  $ goals        : int  1 0 0 1 2 2 4 2 1 3 ...
##  $ xGoals       : num  0.628 0.675 0.876 0.782 0.604 ...
##  $ shots        : int  9 9 11 7 10 11 19 11 17 11 ...
##  $ shotsOnTarget: int  1 4 2 3 5 5 8 5 6 7 ...
##  $ deep         : int  4 10 11 2 5 4 5 6 5 10 ...
##  $ ppda         : num  13.83 8.22 6.9 11.85 6.65 ...
##  $ fouls        : int  12 12 13 13 7 13 13 17 14 20 ...
##  $ corners      : int  1 2 6 3 8 2 6 3 1 4 ...
##  $ yellowCards  : int  2 3 3 4 1 2 2 4 1 0 ...
##  $ redCards     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ result       : chr  "W" "L" "L" "W" ...
```

Gestió de dades invàlides

Per comprovar quines columnes contenen dades 'buides' i poder-hi treballar, utilitzarem la funció colSums, que aplica una funció a totes les columnes d'un dataframe i després aplica una suma.

```
colSums(is.na(games))
```

```
##            gameID          leagueID            season              date
##                 0                 0                 0                 0
##        homeTeamID        awayTeamID        homeGoals         awayGoals
##                 0                 0                 0                 0
##   homeProbability   drawProbability   awayProbability homeGoalsHalfTime
##                 0                 0                 0                 0
## awayGoalsHalfTime             B365H             B365D             B365A
##                 0                 5                 5                 5
##               BWH               BWD               BWA               IWH
##                 3                 3                 3                18
##               IWD               IWA               PSH               PSD
##                18                18                20                20
##               PSA               WHH               WHD               WHA
##                20                 6                 6                 6
##               VCH               VCD               VCA              PSCH
##                 4                 4                 4                 2
##              PSCD              PSCA
##                 2                 2
```

Veiem que hi han un grup de columnes que contenen dades buides. Com que aquestes columnes no ens interesan per a les futures visualitzacions les anem a eliminar.

```
games<- games[1:13]
colSums(is.na(games))
```

```
##            gameID          leagueID            season              date
##                 0                 0                 0                 0
##        homeTeamID        awayTeamID        homeGoals         awayGoals
##                 0                 0                 0                 0
##   homeProbability   drawProbability   awayProbability homeGoalsHalfTime
##                 0                 0                 0                 0
## awayGoalsHalfTime
##                 0
```

```
colSums(is.na(leagues))
```

```
##          leagueID              name understatNotation
##                 0                 0                 0
```

```
colSums(is.na(players))
```

```
## playerID     name
##        0        0
```

```
colSums(is.na(teams))
```

```
## teamID   name
##      0      0
```

```
colSums(is.na(shots))
```

```
##     gameID  shooterID assisterID     minute  situation lastAction   shotType
##          0          0      84344          0          0          0          0
## shotResult      xGoal  positionX  positionY
##          0          0          0          0
```

En aquest cas veiem que tenim a la columna asssiterID un nombre significant de dades buides pero des de el punt de vista futbolistic això és molt normal ja que hi ha xuts que son de jugada individual del jugador en el cual no es requereix cap passador.

```
colSums(is.na(appearances))
```

```
##        gameID      playerID         goals      ownGoals         shots
##             0             0             0             0             0
##        xGoals   xGoalsChain xGoalsBuildup       assists     keyPasses
##             0             0             0             0             0
##      xAssists      position positionOrder    yellowCard       redCard
##             0             0             0             0             0
##          time  substituteIn substituteOut      leagueID
##             0             0             0             0
```

```
colSums(is.na(teamstats))
```

```
##        gameID        teamID        season          date      location
##             0             0             0             0             0
##         goals        xGoals         shots shotsOnTarget          deep
##             0             0             0             0             0
##          ppda         fouls       corners   yellowCards      redCards
##             0             0             0             1             0
##        result
##             0
```

```
#Creem un nou csv file amb les dades games final
write.csv(games, "games_2.csv",  row.names = FALSE)
```

```
 tbls = c("leagues", "players", "teams", "games_2", "shots", "appearances", "teamstats")

for (tbl in tbls) {
    varName = str_c(tbl, "tbl", sep = ".")
   df = read.csv(str_c("/Users/jespinlo10/Documents/Master/2on Semestre/Visualizacion/PRAC1/archive/",
                  stringsAsFactors = FALSE,
                  encoding = "latin1")
   assign(varName, df)
}
```

En total tenim set taules, que inclou informació sobre tots els partits jugats del 2014 al 2020 a les lligues Top 5 d'Europa

```
metaData = tibble()
for (i in 1:length(tbls)) {
  currTbl = get(paste(tbls[i], ".tbl", sep = ""))

  metaData = rbind(metaData, t(c(paste(tbls[i], ".tbl", sep = ""), length(currTbl), nrow(currTbl))))
}
names(metaData) = c("tableName", "variables", "observations")

knitr::kable(metaData)
```

tableName	variables	observations
leagues.tbl	3	5
players.tbl	2	7659
teams.tbl	2	146
games_2.tbl	13	12680
shots.tbl	11	324543
appearances.tbl	19	356513
teamstats.tbl	16	25360

Les lligues i les seves diferències

Com ja s'ha esmentat, la base de dades conté informació sobre les lligues Top 5 d'Europa, incloses la Premier League (Anglaterra), La Liga (Espanya), la Bundesliga (Alemanya), la Sèrie A (Itàlia) i la Ligue 1 (França). Com que ja sabem quants equips participen a cada competició i que cada equip juga amb tots els seus rivals dues vegades (una a casa i una a fora), podem comprovar la integritat de la taula de jocs mirant el nombre de partits per lliga i temporada. Hauríem d'obtenir els següents resultats:

- Premier League: 20 equips i per tant 380 partits
- La Lliga: 20 equips i per tant 380 partits
- Sèrie A: 20 equips i per tant 380 partits
- Bundesliga: 18 equips i per tant 306 partits
- Lliga 1: 20 equips i per tant 380 partits

Tanmateix, després d'agregar el nombre de partits per temporada i lliga i filtrar les quantitats "normals", podem observar alguns valors estranys.

El motiu dels partits perduts de la 2019/20 a la Ligue 1 va ser que el 13 de març de 2020, la LFP (Ligue de Football Professionnel) va suspendre la Ligue 1 indefinidament després de l'esclat de la COVID-19 a França. Totes les altres lligues van continuar jugant després d'un confinament paneuropeu entre març i juny, mentre que els francesos van decidir aturar completament la competició. No obstant això, això no explica l'absència d'aquell partit la 2016/17. El 16 d'abril de 2017, quan l'SC Bastia es va enfrontar amb l'Olympique de Lió, el partit es va suspendre a causa dels seguidors locals, que van envair el terreny de joc dues vegades per atacar als jugadors contraris.

Amb aquest analísis tenim ja els datasets finals per crear visualitzacions que ens donin resposta a las preguntes plantejades a la PRAC1. Per dur a terme la creació de visualitzacions s' utilitzarà public tableau.