R assignment: Football Appearance Dataset

Group 2

2024-06-15

```
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 ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(DescTools)
library(ggplot2)

my_data <- read_csv("Football_dataset.csv")
```

```
## Rows: 1048575 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): appearance_id, date, player_name, competition_id
## dbl (9): game_id, player_id, player_club_id, player_current_club_id, yellow_...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Print the structure of your dataset
str(my_data)
```

```
## spc_tbl_ [1,048,575 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ appearance_id         : chr [1:1048575] "2231978_38004" "2233748_79232" "2234413_42792" "2234418_
##  $ game_id               : num [1:1048575] 2231978 2233748 2234413 2234418 2234421 ...
##  $ player_id             : num [1:1048575] 38004 79232 42792 73333 122011 ...
##  $ player_club_id        : num [1:1048575] 853 8841 6251 1274 195 ...
##  $ player_current_club_id: num [1:1048575] 235 2698 465 6646 3008 ...
##  $ date                  : chr [1:1048575] "03-07-2012" "05-07-2012" "05-07-2012" "05-07-2012" ...
##  $ player_name           : chr [1:1048575] "AurÃ©lien Joachim" "Ruslan Abyshov" "Sander Puri" "Vegar
##  $ competition_id        : chr [1:1048575] "CLQ" "ELQ" "ELQ" "ELQ" ...
##  $ yellow_cards          : num [1:1048575] 0 0 0 0 0 1 0 1 0 0 ...
```

```
##  $ red_cards             : num [1:1048575] 0 0 0 0 0 0 0 0 0 0 ...
##  $ goals                 : num [1:1048575] 2 0 0 0 0 0 0 0 0 0 ...
##  $ assists               : num [1:1048575] 0 0 0 0 1 0 0 1 0 0 ...
##  $ minutes_played        : num [1:1048575] 90 90 45 90 90 90 90 90 45 90 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   appearance_id = col_character(),
##   ..   game_id = col_double(),
##   ..   player_id = col_double(),
##   ..   player_club_id = col_double(),
##   ..   player_current_club_id = col_double(),
##   ..   date = col_character(),
##   ..   player_name = col_character(),
##   ..   competition_id = col_character(),
##   ..   yellow_cards = col_double(),
##   ..   red_cards = col_double(),
##   ..   goals = col_double(),
##   ..   assists = col_double(),
##   ..   minutes_played = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#List the variables in your dataset
variable_list <- names(my_data)
print(variable_list)
```

```
##  [1] "appearance_id"          "game_id"                "player_id"
##  [4] "player_club_id"         "player_current_club_id" "date"
##  [7] "player_name"            "competition_id"         "yellow_cards"
## [10] "red_cards"              "goals"                  "assists"
## [13] "minutes_played"
```

```
#Print the top 15 rows of your dataset
rows_top15 <- head(my_data, n=15)
print(rows_top15)
```

```
## # A tibble: 15 x 13
##    appearance_id  game_id player_id player_club_id player_current_club_id date
##    <chr>            <dbl>     <dbl>          <dbl>                   <dbl> <chr>
##  1 2231978_38004  2231978     38004            853                     235 03-07~
##  2 2233748_79232  2233748     79232           8841                    2698 05-07~
##  3 2234413_42792  2234413     42792           6251                     465 05-07~
##  4 2234418_73333  2234418     73333           1274                    6646 05-07~
##  5 2234421_122011 2234421    122011            195                    3008 05-07~
##  6 2234421_146889 2234421    146889            195                     190 05-07~
##  7 2235539_28716  2235539     28716            282                    7185 05-07~
##  8 2235539_69445  2235539     69445            282                   19771 05-07~
##  9 2235545_19409  2235545     19409            317                     200 05-07~
## 10 2235545_30003  2235545     30003            317                     317 05-07~
## 11 2235545_30667  2235545     30667            317                     317 05-07~
## 12 2235545_34129  2235545     34129            317                    1435 05-07~
## 13 2235545_36139  2235545     36139            317                      36 05-07~
## 14 2235545_4520   2235545      4520            317                     317 05-07~
```

```
## 15 2235545_4582   2235545      4582            317                     317 05-07~
## # i 7 more variables: player_name <chr>, competition_id <chr>,
## #   yellow_cards <dbl>, red_cards <dbl>, goals <dbl>, assists <dbl>,
## #   minutes_played <dbl>
```

```
#Write a user defined function using any of the variables from the data set.
calculate_contribution_points <- function(yellow_cards, red_cards, goals, assists)
  {
    points <- (goals * 3) + (assists * 2) - (yellow_cards * 1) - (red_cards * 3)
    return(points)
  }
# Fetching and storing Frist row values of the dataset
yellow_cards <- my_data[1, "yellow_cards"]
red_cards <- my_data[1, "red_cards"]
goals <- my_data[1, "goals"]
assists <- my_data[1, "assists"]
# Call to function to calculate total points of First row values
total_points <- calculate_contribution_points(yellow_cards, red_cards, goals, assists)
print(paste("Total contribution points for the first player:", total_points))
```

```
## [1] "Total contribution points for the first player: 6"
```

```
#Use data manipulation techniques and filter rows based on any logical criteria that exist in your data
player_name_df <- data.frame(my_data$player_name)

players_with_red_cards <- player_name_df %>% filter(my_data$red_cards >= 1)
players_with_red_cards <- unique(players_with_red_cards)
print(paste("Players with red cards are: ",players_with_red_cards))
```

```
## [1] "Players with red cards are:  c(\"Sergiy Dolganskyi\", \"Claudemir\", \"Cillian Sheridan\", \"Mag
ctor VÃ¡zquez\", \"Gertjan De Mets\", \"GrÃ©gory TadÃ©\", \"Lewis Guy\", \"Artur Tlisov\", \"Christian
nez\", \"Senijad Ibricic\", \"Thulani Serero\", \"Zeljko Brkic\", \"Jonathan Brison\", \"Peter Odemwing
a\", \"Caner Erkin\", \n\"Lucas Mendes\", \"Stijn Wuytens\", \"Brede Hangeland\", \"Ryan Koolwijk\", \"
n\", \"Luca Cigarini\", \"Damiano Zanon\", \"Pantelis Kafes\", \"Jacobo Sanz\", \"Antonio BarragÃ¡n\",
cius\", \n\"Abdoul Wahid Sissoko\", \"Pablo ChavarrÃa\", \"David Ospina\", \"YounÃ¨s Belhanda\", \"Marqu
chel Madera\", \"HÃ©ctor Rodas\", \"Pedro Henrique\", \"Semedo\", \"Rafa LÃ³pez\", \"Thiago Motta\", \n
a\", \"JosÃ© Manuel FernÃ¡ndez\", \"Miguel Villarejo\", \"SÃ©bastien Pocognoli\", \"Maximilian Haas\",
s\", \"MartÃn Demichelis\", \"Felipe Melo\", \"Alberto Aquilani\", \"Manolis Papasterianos\", \"Diego G
n\", \n\"Fernando Navarro\", \"Ramon Zomer\", \"Jeroen Zoet\", \"Ibrahim Ayew\", \"Benjamin Mokulu\", \
zek\", \"Guy Ramos\", \"Theo Janssen\", \"Kassim Abdallah\", \"Shawn Parker\", \n\"Ã"scar Cardozo\", \"
ctor ValdÃ©s\", \"Antonio Candreva\", \"Jonathan Page\", \"Martin Albrechtsen\", \"Francesco Pisano\",
chel Herrero\", \"LuÃs Neto\", \"GÃ¶khan Zan\", \"Jelle Van Damme\", \"John Rankin\", \"Thomas Bruns\",
mer Toprak\", \"GermÃ¡n Denis\", \"SebastiÃ¡n Blanco\", \"Rasmus WÃ¼rtz\", \"Mathieu Flamini\", \"Milan
rez\", \"Danny Fox\", \"Marc-Antoine FortunÃ©\", \"Mauricio Pinilla\", \"David Raven\", \"Nicky Riley\"
ctor SÃ¡nchez\", \"Selcuk Sahin\", \"Volkan Demirel\", \"Sabri Sarioglu\", \"Charlton Vicento\", \"Jason
tor Murta\", \"Ricardo Silva\", \"Paulinho\", \"AkÃ¨s da Costa Goore\", \"Ronan Le Crom\", \"Iker Munia
zbayraktar\", \"Yannick Sagbo\", \n\"Paolo Castellini\", \"Ibrahima TraorÃ©\", \"KrisztiÃ¡n AdorjÃ¡n\",
s\", \"Nikola Aksentijevic\", \"Markus Henriksen\", \"Lass Bangoura\", \"Julian Palmieri\", \"Jamie Ham
ctor Ruiz\", \"Sito Riera\", \"Hrvoje Cale\", \"Musa Nizam\", \"Georgios Ioannidis\", \"Vasilios Koutsia
s Martins\", \"Gino Coutinho\", \"Ã¤mit Kurt\", \"Georgios Dasios\", \"Andreas Tatos\", \"FrÃ©dÃ©ric Fra
to\", \"Mario Yepes\", \"Jeroen Veldmate\", \"Jamal ThiarÃ©\", \"StefÃ¡n GÃslason\", \"Nathan Sinkala\"
ctor Ibarbo\", \"RÃºben Fernandes\", \"KrisztiÃ¡n NÃ©meth\", \"Cedrick\", \"Mamoutou N'Diaye\", \"Jairo
nez\", \"Adam Sarota\", \"VÃctor Ã\\u0081lvarez\", \"GrÃ©gory Lorenzi\", \"Daniel Aranzubia\", \"Filip
```

```
tez\", \"Stipe Perica\", \"Ersan GÃ¼lÃ¼m\", \"Marco Benassi\", \"Daniele Conti\", \"Diego Romano\", \n\"
nz\", \"Kyriakos Papadopoulos\", \"Jovan Kostovski\", \"Lloyd Palun\", \"Mario Sampirisi\", \"VÃ-
ctor CasadesÃºs\", \"Christian TrÃ¤sch\", \"Alexander Fischer\", \"Andreas Cornelius\", \"Adama Soumaoro
tor\", \"Manuel Schmiedebach\", \"Cristian Ledesma\", \"Vladimir Stojkovic\", \"Marcin Komorowski\", \"
a\", \"Papy Djilobodji\", \"Bruno Ecuele Manga\", \"Igor Skoba\", \"Suso\", \"Mikel GonzÃ¡lez\", \"Senao
zyakup\", \"Paolo Cannavaro\", \"Mattia De Sciglio\", \"Ondrej Mazuch\", \"JÃºnior Moraes\", \"Efe Ambro
ncula\", \"Giulio Donati\", \"Loukas Vyntra\", \"Tyler Blackett\", \"Iago Santos\", \n\"VladimÃ-
r Darida\", \"Timothy Derijck\", \"Yannick Cahuzac\", \"Mark Engberink\", \"Emmanuel Boateng\", \"Walid
a Basanta\", \"Dino Arslanagic\", \"Stefan Mitrovic\", \"Ugur Demirok\", \"Dougie Imrie\", \"Danny Hoese
a\", \"FÃ¡bio CoentrÃ£o\", \"Matty James\", \"Henrik Ojamaa\", \"Jander\", \"Anastasios Venetis\", \n\"
zbayrakli\", \"Salim CissÃ©\", \"Ahmet Calik\", \"John Stones\", \"Aymen Abdennour\", \"Marc TorrejÃ³n\
a\", \"Viktor Genev\", \"Adem BÃ¼yÃ¼k\", \"Baris Basdas\", \"Emiliano VelÃ¡zquez\", \"FranÃ§ois Mouband
ztorun\", \"Mark Brown\", \"Vlada Avramov\", \"MaurÃcio\", \"Javier Aquino\", \"Roberto Lago\", \"RaÃºl
zarslan\", \"JoÃ£o Pereira\", \"Vadym Sapay\", \"Aleksandr Sapeta\", \"Jonathan Buatu\", \"John Terry\"
guez\", \"Vincent Enyeama\", \"FÃ¡bio Pacheco\", \"Gabriel Paulista\", \"Kevin Mirallas\", \"Serges Debl
ztÃ¼rk\", \"Gavin Reilly\", \"KakÃ¡\", \"Lamine Gassama\", \"Bogdan Shust\", \"Gabriel Paletta\", \"Baka
s Leal\", \"Andrew Hogg\", \"Ã‰douard Duplan\", \"Anthony Le Tallec\", \"Keita BaldÃ©\", \"Tomi Juric\"
cio AntÃ´nio\", \"Jamiro Monteiro\", \"Jamie Reckord\", \"Nicolas N'Koulou\", \"Rafael Amorim\", \"Paul
a\", \"Cyril ThÃ©rÃ©au\", \"Pawel Wszolek\", \"Michel Morganella\", \"Arghus\", \"Abiola Dauda\", \"Chai
a\", \"Yoel RodrÃguez\", \"Emil Krafth\", \"JosÃ© RodrÃguez\", \"Miguel Veloso\", \"TiemouÃ© Bakayoko\"
zgenÃ§\", \"Mikel Balenziaga\", \"Blerim Dzemaili\", \"Laurens Paulussen\", \"Rami Gershon\", \"Ricardo
n\", \"Valerio Verre\", \"Alex Cordaz\", \"Isaac Mbenza\", \"JoÃ£o PatrÃ£o\", \"Christos Aravidis\", \"
as Nahuel\", \"Deniz Yilmaz\", \"Mathew Leckie\", \"David Abraham\", \"Georgios Saramantas\", \"Dele Ali
lson Junior\", \n\"Vangelis Ikonomou\", \"Ivan Radovanovic\", \"Albano Bizzarri\", \"Jordan Amavi\", \"
s Machado\", \"NoÃ© Acosta\", \"Nikolaos Papadopoulos\", \"Aleksandr Krivoruchko\", \"Rekeem Harper\",
guez\", \"JoÃ£o Gamboa\", \"Taiwo Awoniyi\", \"Rodolfo\", \"Jason Kerr\", \"Anthony Knockaert\", \"Samu
a\", \"Ondrej Celustka\", \"Alejandro GÃ¡lvez\", \"Jozo Simunovic\", \"Adrien TrÃ©bel\", \"Ofir Marciano
colas\", \"Francesco Vicari\", \"Andrew Davies\", \"Abdallah Ndour\", \n\"Claud Adjapong\", \"Darryl Lac
n Correa\", \"JoÃ£o GÃ³is\", \"Victor Andrade\", \"Michiel Kramer\", \"Mattia Bani\", \"Vincent Sasso\"
n Rolle\", \"Marcel Heller\", \"Nabil Bentaleb\", \"Marc Albrighton\", \n\"Juan Jesus\", \"MatÃ-
as Vecino\", \"Gabriel Pires\", \"RubÃ©n Pardo\", \"Adama Diakhaby\", \"Dusko Tosic\", \"Platiny\", \"Ar
a\", \"Bakary KonÃ©\", \n\"Miiko Albornoz\", \"JoÃ£o Lucas\", \"Matic Fink\", \"Abdoulaye Ba\", \"IvÃ¡n
n\", \"Igor Plastun\", \"Josip Ilicic\", \"Yevgen Khacheridi\", \"Merih Demiral\", \"Steve MouniÃ©\", \
rez\", \"Jesper Drost\", \"Lennart Czyborra\", \"JuliÃ¡n Cuesta\", \"Aleksandr Golovin\", \"SadÄ±k Ã‡if
zer\", \"Josip Vukovic\", \"Jorrit Smeets\", \"Yuki Kobayashi\", \"Ben Rienstra\", \n\"Vitor Hugo\", \"
a\", \"Alexandre Lacazette\", \"Suat Serdar\", \"Adis Jahovic\", \"Shaun Byrne\", \"Ryad Boudebouz\", \
bar\", \"Pedro Trigueira\", \"Renan Ribeiro\", \"Nicolai JÃ¶rgensen\", \"Mikhail Merkulov\", \"Juan Berr
guez\", \"Jorge Molina\", \"Joel Robles\", \"Fabinho\", \"Jamie Hamilton\", \"Keven Schlotterbeck\", \"
ztÃ¼rk\", \"Nazar Verbnyi\", \n\"Rafael Sabino\", \"Menno Koch\", \"Juan MuÃ±iz\", \"Bjarke Jacobsen\",
ctor Laguardia\", \"Lorenzo Burnet\", \"Daniel Anyembe\", \"Joey Konings\", \"Karim Hafez\", \"Oleg Kudi
nez\", \"Federico Peluso\", \"JoÃ£o Novais\", \"Gelson Martins\", \"Dimitry Bertaud\", \"Gerrit Holtmann
az\", \"Vasilios Mantzis\", \"Marcel Langer\", \"Farid Boulaya\", \"Ruben Gabrielsen\", \"Casimir Ninga\
n Aguirregabiria\", \"Guram Tetrashvili\", \"Majid Hosseini\", \"Lucas Sasha\", \"Amir Hadziahmetovic\"
lvio\", \"Roman Zobnin\", \"Marco D'Alessandro\", \"Vladimir Poluyakhtov\", \"Andrea Carboni\", \"Eddie
```

```r
print(paste("Total number of players with red cards: ", nrow(players_with_red_cards)))
```

```
## [1] "Total number of players with red cards:  3043"
```

```r
#Identify the dependent & independent variables and use reshaping techniques and create a new data fram
appearance_id<-my_data$appearance_id
player_id<-my_data$player_id
player_name<-my_data$player_name
```

```r
player_club_id<-my_data$player_club_id
game_id<-my_data$game_id
competition_id<-my_data$competition_id

player_df <- data.frame(appearance_id,player_id,player_name,player_club_id)
game_df <- data.frame(appearance_id,game_id,competition_id)

merged_df <- merge(player_df,game_df,by="appearance_id")
head(merged_df)
```

```
##    appearance_id player_id       player_name player_club_id game_id
## 1 2211607_111184    111184      Dico Koppers            610 2211607
## 2  2211607_12282     12282       Daley Blind            610 2211607
## 3 2211607_124883    124883 Ricardo van Rhijn            610 2211607
## 4 2211607_124891    124891     Aras Ã-zbiliz            610 2211607
## 5 2211607_146258    146258     Jetro Willems            383 2211607
## 6  2211607_16101     16101  Atiba Hutchinson            383 2211607
##   competition_id
## 1           NLSC
## 2           NLSC
## 3           NLSC
## 4           NLSC
## 5           NLSC
## 6           NLSC
```

```r
#Remove missing values in your dataset.
my_data <- my_data %>% filter(!is.na(my_data$appearance_id),!is.na(my_data$game_id),!is.na(my_data$play
head(my_data)
```

```
## # A tibble: 6 x 13
##   appearance_id  game_id player_id player_club_id player_current_club_id date
##   <chr>            <dbl>     <dbl>          <dbl>                  <dbl> <chr>
## 1 2231978_38004  2231978     38004            853                    235 03-07-~
## 2 2233748_79232  2233748     79232           8841                   2698 05-07-~
## 3 2234413_42792  2234413     42792           6251                    465 05-07-~
## 4 2234418_73333  2234418     73333           1274                   6646 05-07-~
## 5 2234421_122011 2234421    122011            195                   3008 05-07-~
## 6 2234421_146889 2234421    146889            195                    190 05-07-~
## # i 7 more variables: player_name <chr>, competition_id <chr>,
## #   yellow_cards <dbl>, red_cards <dbl>, goals <dbl>, assists <dbl>,
## #   minutes_played <dbl>
```

```r
#Identify and remove duplicated data in your dataset
my_data <- my_data %>% distinct()
head(my_data)
```

```
## # A tibble: 6 x 13
##   appearance_id  game_id player_id player_club_id player_current_club_id date
##   <chr>            <dbl>     <dbl>          <dbl>                  <dbl> <chr>
## 1 2231978_38004  2231978     38004            853                    235 03-07-~
## 2 2233748_79232  2233748     79232           8841                   2698 05-07-~
## 3 2234413_42792  2234413     42792           6251                    465 05-07-~
```

```
## 4 2234418_73333  2234418     73333           1274                    6646 05-07-~
## 5 2234421_122011 2234421    122011            195                    3008 05-07-~
## 6 2234421_146889 2234421    146889            195                     190 05-07-~
## # i 7 more variables: player_name <chr>, competition_id <chr>,
## #   yellow_cards <dbl>, red_cards <dbl>, goals <dbl>, assists <dbl>,
## #   minutes_played <dbl>
```

```
#Reorder multiple rows in descending order
my_data <- my_data %>% arrange(desc(player_id),desc(goals))
head(my_data)
```

```
## # A tibble: 6 x 13
##   appearance_id  game_id player_id player_club_id player_current_club_id date
##   <chr>            <dbl>     <dbl>          <dbl>                  <dbl> <chr>
## 1 3251812_830225 3251812    830225          18303                  18303 25-09-~
## 2 3394609_814725 3394609    814725           1160                   1160 27-09-~
## 3 3394587_804934 3394587    804934           2969                   2969 13-09-~
## 4 3394591_804934 3394591    804934           2969                   2969 20-09-~
## 5 3393772_797358 3393772    797358           2999                   2999 08-08-~
## 6 3394869_797358 3394869    797358           2999                   2999 26-09-~
## # i 7 more variables: player_name <chr>, competition_id <chr>,
## #   yellow_cards <dbl>, red_cards <dbl>, goals <dbl>, assists <dbl>,
## #   minutes_played <dbl>
```

```
#Rename some of the column names in your dataset
my_data_updated<- my_data %>% rename(player_full_name=player_name, goals_scored=goals, minutes=minutes_
print(paste("Column names after update:"))
```

```
## [1] "Column names after update:"
```

```
print(colnames(my_data_updated))
```

```
##  [1] "appearance_id"          "game_id"                "player_id"
##  [4] "player_club_id"         "player_current_club_id" "date"
##  [7] "player_full_name"       "competition_id"         "yellow_cards"
## [10] "red_cards"              "goals_scored"           "assists"
## [13] "minutes"
```

```
#Add new variables in your data frame by using a mathematical function
my_data<- my_data%>% mutate(contribution_points = goals + assists + 0.5 * yellow_cards - red_cards)
head(my_data$contribution_points)
```

```
## [1] 0.5 0.0 0.0 0.0 0.0 0.0
```

```
#Create a training set using random number generator engine.
set.seed(123)
train_indices <- sample(1:nrow(my_data), 0.7 * nrow(my_data))
train_data <- my_data[train_indices, ]
head(train_data)
```

```
## # A tibble: 6 x 14
##   appearance_id  game_id player_id player_club_id player_current_club_id date
##   <chr>            <dbl>     <dbl>          <dbl>                  <dbl> <chr>
## 1 2335174_15102  2335174     15102           3725                   1186 25-11-~
## 2 2594765_221322 2594765    221322           1049                   1049 20-03-~
## 3 2495307_258626 2495307    258626            294                    114 01-10-~
## 4 3047608_266359 3047608    266359            273                    273 02-12-~
## 5 2250458_56416  2250458     56416            383                    141 05-05-~
## 6 2604328_197747 2604328    197747              5                    398 01-05-~
## # i 8 more variables: player_name <chr>, competition_id <chr>,
## #   yellow_cards <dbl>, red_cards <dbl>, goals <dbl>, assists <dbl>,
## #   minutes_played <dbl>, contribution_points <dbl>
```

```
#Print the summary statistics of your dataset
summary(my_data)
```

```
##  appearance_id         game_id          player_id      player_club_id
##  Length:1048568     Min.   :2211607   Min.   :    10   Min.   :    1
##  Class :character   1st Qu.:2455176   1st Qu.: 42539   1st Qu.:  281
##  Mode  :character   Median :2697623   Median : 85706   Median :  855
##                     Mean   :2711343   Mean   :124943   Mean   : 2644
##                     3rd Qu.:3047704   3rd Qu.:182913   3rd Qu.: 2425
##                     Max.   :3951263   Max.   :830225   Max.   :75635
##  player_current_club_id     date           player_name
##  Min.   :    3          Length:1048568     Length:1048568
##  1st Qu.:  347          Class :character   Class :character
##  Median :  964          Mode  :character   Mode  :character
##  Mean   : 3578
##  3rd Qu.: 2700
##  Max.   :83678
##  competition_id      yellow_cards      red_cards            goals
##  Length:1048568     Min.   :0.0000   Min.   :0.000000   Min.   :0.0000
##  Class :character   1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.0000
##  Mode  :character   Median :0.0000   Median :0.000000   Median :0.0000
##                     Mean   :0.1534   Mean   :0.003868   Mean   :0.0982
##                     3rd Qu.:0.0000   3rd Qu.:0.000000   3rd Qu.:0.0000
##                     Max.   :2.0000   Max.   :1.000000   Max.   :6.0000
##     assists        minutes_played  contribution_points
##  Min.   :0.00000   Min.   :  1.0   Min.   :-1.000
##  1st Qu.:0.00000   1st Qu.: 60.0   1st Qu.: 0.000
##  Median :0.00000   Median : 90.0   Median : 0.000
##  Mean   :0.07801   Mean   : 71.4   Mean   : 0.249
##  3rd Qu.:0.00000   3rd Qu.: 90.0   3rd Qu.: 0.500
##  Max.   :6.00000   Max.   :135.0   Max.   : 8.000
```

```
#Use any of the numerical variables from the dataset and perform the following statistical functions: M
mean_value<- mean(my_data$minutes_played)
print(paste("Mean value: ",mean_value))
```

```
## [1] "Mean value:  71.3984987144372"
```

```r
median_value<-median(my_data$minutes_played)
print(paste("Median value: ",median_value))
```

```
## [1] "Median value:  90"
```

```r
mode_value <- Mode(my_data$minutes_played)
print(paste("Mode value: ",mode_value))
```

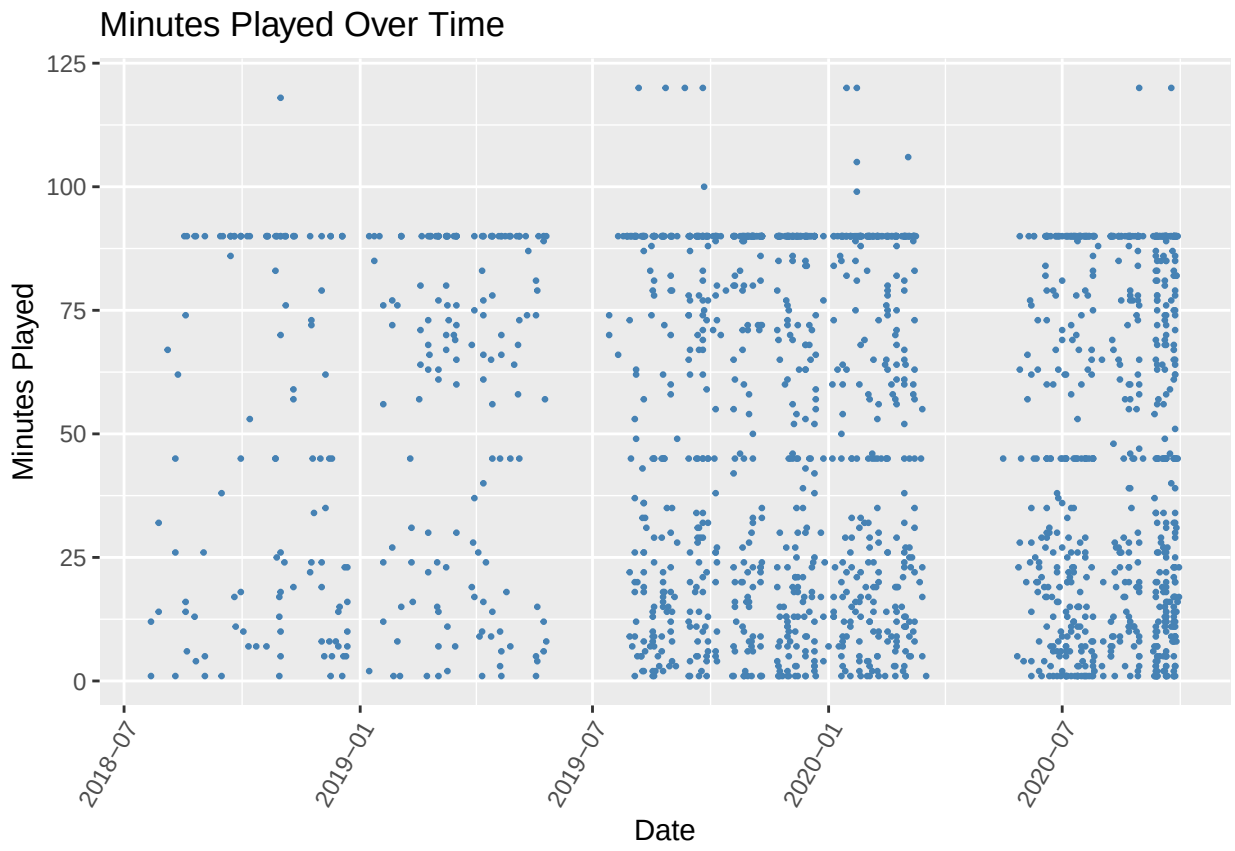
```
## [1] "Mode value:  90"
```

```r
range_value<-range(my_data$minutes_played)
print(paste("Range value: ",range_value))
```

```
## [1] "Range value:  1"   "Range value:  135"
```

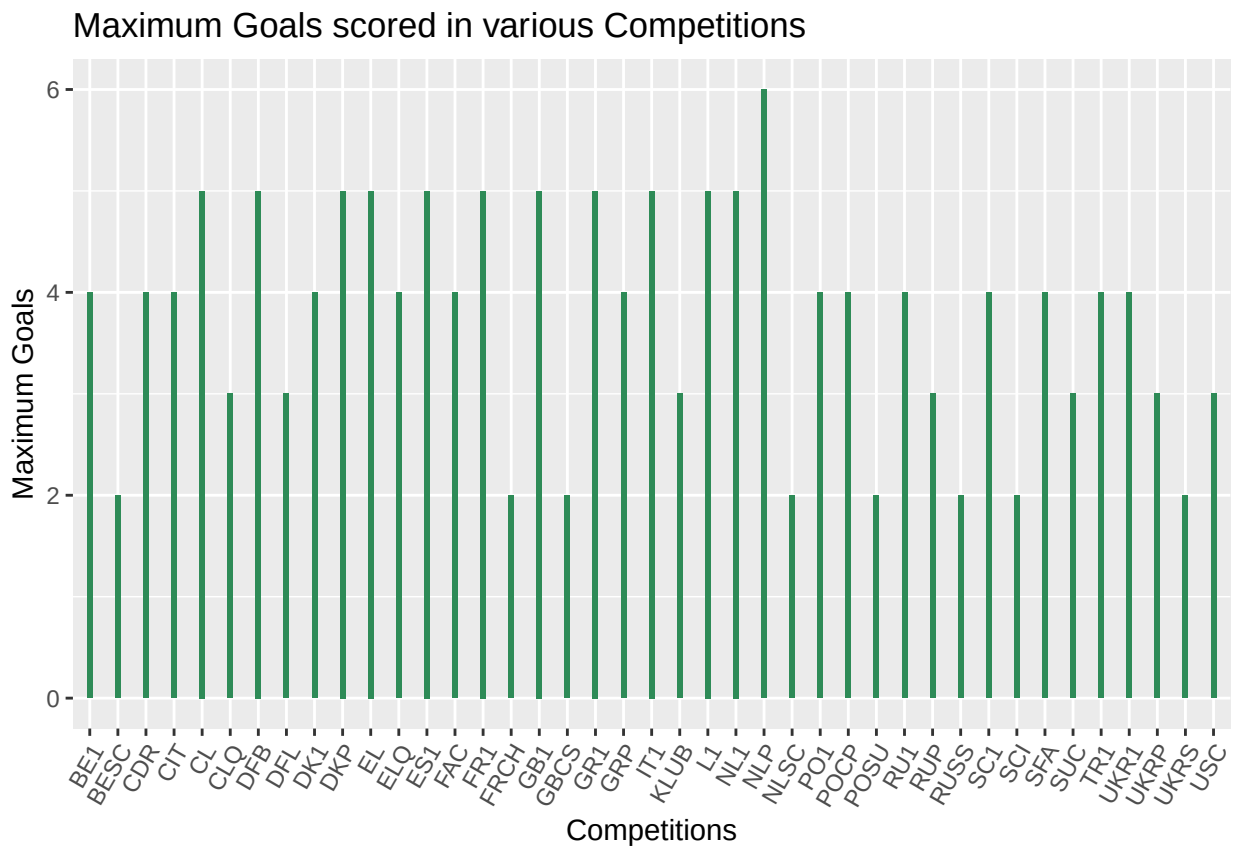
```r
#Plot a scatter plot for any 2 variables in your dataset

my_data_subset <- my_data[1:2000, ]
my_data_subset$date <- as.Date(my_data_subset$date, format = "%d-%m-%Y")
ggplot(my_data_subset, aes(x = date, y = minutes_played, color = goals)) + geom_point(color = "steelblu
labs(title = "Minutes Played Over Time",x = "Date",y = "Minutes Played",color="Goals") +
theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



```r
#Plot a bar plot for any 2 variables in your dataset
new_my_data <- my_data %>% group_by(competition_id) %>% summarise(max_goals = max(goals))

ggplot(new_my_data, aes(x = as.factor(competition_id), y = max_goals)) +
geom_bar(stat = "identity", fill = "seagreen4", width = 0.2) +
labs(title = "Maximum Goals scored in various Competitions",x = "Competitions",y = "Maximum Goals") +
theme(axis.text.x =element_text(angle=60, hjust=1))
```




```r
#Find the correlation between any 2 variables by applying Pearson correlation
correlation <- cor(my_data$yellow_cards, my_data$minutes_played, method = "pearson")
print(correlation)
```

```
## [1] 0.1081577
```