EuroCup Soccer

Shu Liu

2023-06-02

EuroCup Soccer

This is Data in Motion data analysis challenge #2 More details click here link

Scenario

You are a sports data analyst and you have been tasked with summarizing data from the matches from a previous EuroCup. Your manager would like the following questions answered.

Get the data

Download dataset: Link to dataset

Challenge Questions

1. How many teams participated in the Euro2012?
2. What is the number of columns in the dataset?
3. View only the columns Team, Yellow Cards and Red Cards and assign them to a dataframe called discipline.
4. Sort the teams by Red Cards, then to Yellow Cards.
5. Calculate the mean Yellow Cards given per Team.
6. Filter teams that scored more than 6 goals.
7. Select the teams that start with the letter G.
8. Select the first 7 columns.
9. Select all columns except the last 3.
10. Present only the Shooting Accuracy from England, Italy and Russia.

Steps

Set up environments

Notes: install package "tidyverse" and "dplyr" for sorting and select.

```
#an argument repos is added to the function that gives it the web address of the repository.
install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\liuch\AppData\Local\Temp\RtmpO0zrs2\downloaded_packages
```

```
library(tidyverse)
library(dplyr) #for sorting and select
```

Load data

Save the data set into local file directory. change working directory to where the file is. load the data into data frame eurocup_soccer.

```
setwd("C:/Users/liuch/OneDrive/Documents/DataAnalytics/Portfolio/case_study_2")
eurocup_soccer <- read_csv("Euro_2012_stats_TEAM.csv")
```

Now the eurocup_soccer has the data. Let's take a glimpse.

```
glimpse(eurocup_soccer)
```

```
## Rows: 16
## Columns: 35
## $ Team                          <chr> "Croatia", "Czech Republic", "Denmark", "~
## $ Goals                         <dbl> 4, 4, 4, 5, 3, 10, 5, 6, 2, 2, 6, 1, 5, 1~
## $ `Shots on target`             <dbl> 13, 13, 10, 11, 22, 32, 8, 34, 12, 15, 22~
## $ `Shots off target`            <dbl> 12, 18, 10, 18, 24, 32, 18, 45, 36, 23, 4~
## $ `Shooting Accuracy`           <chr> "51.9%", "41.9%", "50.0%", "50.0%", "37.9~
## $ `% Goals-to-shots`            <chr> "16.0%", "12.9%", "20.0%", "17.2%", "6.5%~
## $ `Total shots (inc. Blocked)` <dbl> 32, 39, 27, 40, 65, 80, 32, 110, 60, 48, ~
## $ `Hit Woodwork`                <dbl> 0, 0, 1, 0, 1, 2, 1, 2, 2, 0, 6, 0, 2, 0,~
## $ `Penalty goals`               <dbl> 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1,~
## $ `Penalties not scored`        <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ `Headed goals`                <dbl> 2, 0, 3, 3, 0, 2, 0, 2, 0, 1, 2, 1, 1, 2,~
## $ Passes                        <dbl> 1076, 1565, 1298, 1488, 2066, 2774, 1187,~
## $ `Passes completed`            <dbl> 828, 1223, 1082, 1200, 1803, 2427, 911, 2~
## $ `Passing Accuracy`            <chr> "76.9%", "78.1%", "83.3%", "80.6%", "87.2~
## $ Touches                       <dbl> 1706, 2358, 1873, 2440, 2909, 3761, 2016,~
## $ Crosses                       <dbl> 60, 46, 43, 58, 55, 101, 52, 75, 50, 55, ~
## $ Dribbles                      <dbl> 42, 68, 32, 60, 76, 60, 53, 75, 49, 39, 6~
## $ `Corners Taken`               <dbl> 14, 21, 16, 16, 28, 35, 10, 30, 22, 14, 4~
## $ Tackles                       <dbl> 49, 62, 40, 86, 71, 91, 65, 98, 34, 67, 7~
## $ Clearances                    <dbl> 83, 98, 61, 106, 76, 73, 123, 137, 41, 87~
## $ Interceptions                 <dbl> 56, 37, 59, 72, 58, 69, 87, 136, 41, 62, ~
## $ `Clearances off line`         <dbl> NA, 2, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0~
## $ `Clean Sheets`                <dbl> 0, 1, 1, 2, 1, 1, 1, 2, 0, 0, 2, 0, 0, 5,~
## $ Blocks                        <dbl> 10, 10, 10, 29, 7, 11, 23, 18, 9, 8, 11, ~
## $ `Goals conceded`              <dbl> 3, 6, 5, 3, 5, 6, 7, 7, 5, 3, 4, 9, 3, 1,~
## $ `Saves made`                  <dbl> 13, 9, 10, 22, 6, 10, 13, 20, 12, 6, 10, ~
## $ `Saves-to-shots ratio`        <chr> "81.3%", "60.1%", "66.7%", "88.1%", "54.6~
## $ `Fouls Won`                   <dbl> 41, 53, 25, 43, 36, 63, 67, 101, 35, 48, ~
## $ `Fouls Conceded`              <dbl> 62, 73, 38, 45, 51, 49, 48, 89, 30, 56, 9~
## $ Offsides                      <dbl> 2, 8, 8, 6, 5, 12, 12, 16, 3, 3, 10, 11, ~
## $ `Yellow Cards`                <dbl> 9, 7, 4, 5, 6, 4, 9, 16, 5, 7, 12, 6, 6, ~
## $ `Red Cards`                   <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0,~
## $ `Subs on`                     <dbl> 9, 11, 7, 11, 11, 15, 12, 18, 7, 7, 14, 1~
## $ `Subs off`                    <dbl> 9, 11, 7, 11, 11, 15, 12, 18, 7, 7, 14, 1~
## $ `Players Used`                <dbl> 16, 19, 15, 16, 19, 17, 20, 19, 15, 17, 1~
```

```
#check is there any missing values.I found 1 missing value.
sum(is.na(eurocup_soccer))
```

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

Analyze data and answer questions

1. How many teams participated in the Euro2012? from the glimpse() above we can find out there are **16** teams participated in the Euro2012.
2. What is the number of columns in the dataset? From the glimpse() above we can find out there are **35** columns are in the dataset
3. View only the columns Team, Yellow Cards and Red Cards and assign them to a dataframe called discipline.

```
discipline <- eurocup_soccer %>% select(Team,`Yellow Cards`,`Red Cards`)
```

4. Sort the teams by Red Cards, then to Yellow Cards.

```
discipline %>% arrange(`Yellow Cards`, `Red Cards`)
```

```
## # A tibble: 16 x 3
##    Team                ‘Yellow Cards‘ ‘Red Cards‘
##    <chr>                        <dbl>       <dbl>
##  1 Denmark                          4           0
##  2 Germany                          4           0
##  3 England                          5           0
##  4 Netherlands                      5           0
##  5 Ukraine                          5           0
##  6 France                           6           0
##  7 Russia                           6           0
##  8 Republic of Ireland              6           1
##  9 Czech Republic                   7           0
## 10 Sweden                           7           0
## 11 Poland                           7           1
## 12 Croatia                          9           0
## 13 Greece                           9           1
## 14 Spain                           11           0
## 15 Portugal                        12           0
## 16 Italy                           16           0
```

5. Calculate the mean Yellow Cards given per Team.

```
discipline %>% select(Team, `Yellow Cards`, `Red Cards`) %>% summarise(mean_yellor_cards=mean(`Yellow Ca
```

```
## # A tibble: 1 x 1
##   mean_yellor_cards
##               <dbl>
## 1              7.44
```

The mean Yellow Cards given per Team is **7.44**

6. Filter teams that scored more than 6 goals.

```
eurocup_soccer %>% filter(Goals>6)
```

```
## # A tibble: 2 x 35
##   Team    Goals `Shots on target` `Shots off target` `Shooting Accuracy`
##   <chr>   <dbl>             <dbl>              <dbl> <chr>
## 1 Germany    10                32                 32 47.8%
## 2 Spain      12                42                 33 55.9%
## # i 30 more variables: `% Goals-to-shots` <chr>,
## #   `Total shots (inc. Blocked)` <dbl>, `Hit Woodwork` <dbl>,
## #   `Penalty goals` <dbl>, `Penalties not scored` <dbl>, `Headed goals` <dbl>,
## #   Passes <dbl>, `Passes completed` <dbl>, `Passing Accuracy` <chr>,
## #   Touches <dbl>, Crosses <dbl>, Dribbles <dbl>, `Corners Taken` <dbl>,
## #   Tackles <dbl>, Clearances <dbl>, Interceptions <dbl>,
## #   `Clearances off line` <dbl>, `Clean Sheets` <dbl>, Blocks <dbl>, ...
```

Only Germany and Spain have more than 6 goals.

7. Select the teams that start with the letter G. We can use filter and subtr function

```
eurocup_soccer %>% filter(substr(Team,1,1)=="G")
```

```
## # A tibble: 2 x 35
##   Team    Goals `Shots on target` `Shots off target` `Shooting Accuracy`
##   <chr>   <dbl>             <dbl>              <dbl> <chr>
## 1 Germany    10                32                 32 47.8%
## 2 Greece      5                 8                 18 30.7%
## # i 30 more variables: `% Goals-to-shots` <chr>,
## #   `Total shots (inc. Blocked)` <dbl>, `Hit Woodwork` <dbl>,
## #   `Penalty goals` <dbl>, `Penalties not scored` <dbl>, `Headed goals` <dbl>,
## #   Passes <dbl>, `Passes completed` <dbl>, `Passing Accuracy` <chr>,
## #   Touches <dbl>, Crosses <dbl>, Dribbles <dbl>, `Corners Taken` <dbl>,
## #   Tackles <dbl>, Clearances <dbl>, Interceptions <dbl>,
## #   `Clearances off line` <dbl>, `Clean Sheets` <dbl>, Blocks <dbl>, ...
```

Only Germany and Greece start with G.

8. Select the first 7 columns.

```
eurocup_soccer %>% select(1:7)
```

```
## # A tibble: 16 x 7
##    Team           Goals `Shots on target` `Shots off target` `Shooting Accuracy`
##    <chr>          <dbl>             <dbl>              <dbl> <chr>
##  1 Croatia            4                13                 12 51.9%
##  2 Czech Republic     4                13                 18 41.9%
##  3 Denmark            4                10                 10 50.0%
```

```
##  4 England            5                11                 18 50.0%
##  5 France             3                22                 24 37.9%
##  6 Germany           10                32                 32 47.8%
##  7 Greece             5                 8                 18 30.7%
##  8 Italy              6                34                 45 43.0%
##  9 Netherlands        2                12                 36 25.0%
## 10 Poland             2                15                 23 39.4%
## 11 Portugal           6                22                 42 34.3%
## 12 Republic of I~     1                 7                 12 36.8%
## 13 Russia             5                 9                 31 22.5%
## 14 Spain             12                42                 33 55.9%
## 15 Sweden             5                17                 19 47.2%
## 16 Ukraine            2                 7                 26 21.2%
## # i 2 more variables: `% Goals-to-shots` <chr>,
## #   `Total shots (inc. Blocked)` <dbl>
```

9. Select all columns except the last 3.

```
eurocup_soccer %>% select(1:(ncol(eurocup_soccer) - 3))
```

```
## # A tibble: 16 x 32
##    Team           Goals `Shots on target` `Shots off target` `Shooting Accuracy`
##    <chr>          <dbl>             <dbl>               <dbl> <chr>
##  1 Croatia            4                13                  12 51.9%
##  2 Czech Republic     4                13                  18 41.9%
##  3 Denmark            4                10                  10 50.0%
##  4 England            5                11                  18 50.0%
##  5 France             3                22                  24 37.9%
##  6 Germany           10                32                  32 47.8%
##  7 Greece             5                 8                  18 30.7%
##  8 Italy              6                34                  45 43.0%
##  9 Netherlands        2                12                  36 25.0%
## 10 Poland             2                15                  23 39.4%
## 11 Portugal           6                22                  42 34.3%
## 12 Republic of I~     1                 7                  12 36.8%
## 13 Russia             5                 9                  31 22.5%
## 14 Spain             12                42                  33 55.9%
## 15 Sweden             5                17                  19 47.2%
## 16 Ukraine            2                 7                  26 21.2%
## # i 27 more variables: `% Goals-to-shots` <chr>,
## #   `Total shots (inc. Blocked)` <dbl>, `Hit Woodwork` <dbl>,
## #   `Penalty goals` <dbl>, `Penalties not scored` <dbl>, `Headed goals` <dbl>,
## #   Passes <dbl>, `Passes completed` <dbl>, `Passing Accuracy` <chr>,
## #   Touches <dbl>, Crosses <dbl>, Dribbles <dbl>, `Corners Taken` <dbl>,
## #   Tackles <dbl>, Clearances <dbl>, Interceptions <dbl>,
## #   `Clearances off line` <dbl>, `Clean Sheets` <dbl>, Blocks <dbl>, ...
```

10. Present only the Shooting Accuracy from England, Italy and Russia.

```
teams_interests<-c("England","Italy","Russia") ## create a vector
## create a subdataset while Team in the vector teams_interests
subdata_set <- eurocup_soccer %>% filter(Team %in% teams_interests)
##only select Team and Shooting Accuracy.
subdata_set %>% select(Team, `Shooting Accuracy`)
```

```
## # A tibble: 3 x 2
##   Team    `Shooting Accuracy`
##   <chr>   <chr>
## 1 England 50.0%
## 2 Italy   43.0%
## 3 Russia  22.5%
```

Notes. This dataset is small and clean. So it's easy to analyze. I also notice if there's space in the column name when we need to use it we need to add **backtitks ‘** to quota it, such as **‘Yellow Cards‘**.