Laliga dataset

Khanh Le

2024-05-31

Import The Dataset ()

```
pacman::p_load(
  ggplot2,
  mvtnorm,
  GGally,
  corrplot,
  readxl,
  tidyverse,
  gridExtra,
  grid,
  plotly,
  ggcorrplot,
  FactoMineR,
  factoextra
)
```

```
data = read_excel("la_liga.xlsx")
```

Data and Methods ():

Information About The Dataset

- `Team` : Name of the football team.
- `Points` : Number of points earned in the season.
- `Matches` : Total number of matches played in the season.
- `Wins` : Number of matches won.
- `Draws` : Number of matches drawn.
- `Loses` : Number of matches lost.
- `Goals.scored` : Number of goals scored.
- `Goals.conceded` : Number of goals conceded.
- `Difference.goals` : Goal difference (goals scored minus goals conceded).
- `Percentage.scored.goals` : Average number of goals scored per match.
- `Percentage.conceded.goals` : Average number of goals conceded per match.
- `Shots` : Total number of shots.
- `Shots.on.goal` : Number of shots on target.
- `Penalties.scored` : Number of goals scored from penalties.
- `Assistances` : Number of assists.
- `Fouls.made` : Number of fouls committed.
- `Matches.without.conceding` : Number of matches with a clean sheet (no goals conceded).
- `Yellow.cards` : Number of yellow cards received.
- `Red.cards` : Number of red cards received.
- `Offsides` : Number of offsides.

Remark on the dataset:

The dataset contains 20 rows and 20 columns which no duplicate or missing values. All columns are positive number.

Methods:

To identify the indicators that influence football performance, we perform a comparative analysis between teams of varying success levels. Often, we encounter datasets with many related categories; hence, applying techniques to reduce the quantity of data can be beneficial. In this study, we aim to reduce the dimensions of a data matrix without losing relevant information by using Principal Component Analysis (PCA). Subsequently, we utilize these principal components to identify the performance differences between the top and bottom teams in LaLiga.

```
head(data)
```

```
## # A tibble: 6 × 20
##   Team            Points Matches  Wins Draws Loses Goals.scored Goals.conceded
##   <chr>            <dbl>   <dbl> <dbl> <dbl> <dbl>        <dbl>          <dbl>
## 1 Barcelona           91      38    29     4     5          112             29
## 2 Real Madrid         90      38    28     6     4          110             34
## 3 Atlético Madrid     88      38    28     4     6           63             18
## 4 Villarreal          64      38    18    10    10           44             35
## 5 Athletic            62      38    18     8    12           58             45
## 6 Celta               60      38    17     9    12           51             59
## # ℹ 12 more variables: Difference.goals <dbl>, Percentage.scored.goals <dbl>,
## #   Percentage.conceded.goals <dbl>, Shots <dbl>, Shots.on.goal <dbl>,
## #   Penalties.scored <dbl>, Assistances <dbl>, Fouls.made <dbl>,
## #   Matches.without.conceding <dbl>, Yellow.cards <dbl>, Red.cards <dbl>,
## #   Offsides <dbl>
```

```
str(data)
```

```
## tibble [20 × 20] (S3: tbl_df/tbl/data.frame)
##  $ Team                     : chr [1:20] "Barcelona" "Real Madrid" "Atlético Madrid" "Villarreal" ...
##  $ Points                   : num [1:20] 91 90 88 64 62 60 52 48 48 45 ...
##  $ Matches                  : num [1:20] 38 38 38 38 38 38 38 38 38 38 ...
##  $ Wins                     : num [1:20] 29 28 28 18 18 17 14 12 13 11 ...
##  $ Draws                    : num [1:20] 4 6 4 10 8 9 10 12 9 12 ...
##  $ Loses                    : num [1:20] 5 4 6 10 12 12 14 14 16 15 ...
##  $ Goals.scored             : num [1:20] 112 110 63 44 58 51 51 38 45 34 ...
##  $ Goals.conceded           : num [1:20] 29 34 18 35 45 59 50 35 48 52 ...
##  $ Difference.goals         : num [1:20] 83 76 45 9 13 -8 1 3 -3 -18 ...
##  $ Percentage.scored.goals  : num [1:20] 2.95 2.89 1.66 1.16 1.53 1.34 1.34 1 1.18 0.89 ...
##  $ Percentage.conceded.goals: num [1:20] 0.76 0.89 0.47 0.92 1.18 1.55 1.32 0.92 1.26 1.37 ...
##  $ Shots                    : num [1:20] 600 712 481 346 450 442 460 452 454 398 ...
##  $ Shots.on.goal            : num [1:20] 277 299 186 135 178 170 189 170 164 132 ...
##  $ Penalties.scored         : num [1:20] 11 6 1 3 3 4 6 2 1 3 ...
##  $ Assistances              : num [1:20] 79 90 49 32 42 43 35 27 33 26 ...
##  $ Fouls.made               : num [1:20] 385 420 503 534 502 528 555 552 465 490 ...
##  $ Matches.without.conceding: num [1:20] 18 14 24 17 13 10 11 12 13 12 ...
##  $ Yellow.cards             : num [1:20] 66 72 91 100 84 116 106 110 108 110 ...
##  $ Red.cards                : num [1:20] 1 5 3 4 5 6 7 5 5 3 ...
##  $ Offsides                 : num [1:20] 120 114 84 106 92 103 106 85 85 80 ...
```

Exploratory Data Analysis ()

1. Data Overview ()

We will add some features to the original dataset to facilitate data exploration, simultaneously, deviding the teams into two group.

```
data <- data %>%
  mutate(
         Effectiveness.Percentage  = round(((Goals.scored + Shots.on.goal) / Shots) * 100, digits=2),
         Rank = row_number(),
         Team.Level = case_when(
            Rank <= 6 ~ "Top",
            Rank > 6 & Rank < 14 ~ "Middle",
            TRUE ~ "Bottom")
        )
data = data[, -c(22)]
str(data)
```

```
## tibble [20 × 22] (S3: tbl_df/tbl/data.frame)
##  $ Team                     : chr [1:20] "Barcelona" "Real Madrid" "Atlético Madrid" "Villarreal" ...
##  $ Points                   : num [1:20] 91 90 88 64 62 60 52 48 48 45 ...
##  $ Matches                  : num [1:20] 38 38 38 38 38 38 38 38 38 38 ...
##  $ Wins                     : num [1:20] 29 28 28 18 18 17 14 12 13 11 ...
##  $ Draws                    : num [1:20] 4 6 4 10 8 9 10 12 9 12 ...
##  $ Loses                    : num [1:20] 5 4 6 10 12 12 14 14 16 15 ...
##  $ Goals.scored             : num [1:20] 112 110 63 44 58 51 51 38 45 34 ...
##  $ Goals.conceded           : num [1:20] 29 34 18 35 45 59 50 35 48 52 ...
##  $ Difference.goals         : num [1:20] 83 76 45 9 13 -8 1 3 -3 -18 ...
##  $ Percentage.scored.goals  : num [1:20] 2.95 2.89 1.66 1.16 1.53 1.34 1.34 1 1.18 0.89 ...
##  $ Percentage.conceded.goals: num [1:20] 0.76 0.89 0.47 0.92 1.18 1.55 1.32 0.92 1.26 1.37 ...
##  $ Shots                    : num [1:20] 600 712 481 346 450 442 460 452 454 398 ...
##  $ Shots.on.goal            : num [1:20] 277 299 186 135 178 170 189 170 164 132 ...
##  $ Penalties.scored         : num [1:20] 11 6 1 3 3 4 6 2 1 3 ...
##  $ Assistances              : num [1:20] 79 90 49 32 42 43 35 27 33 26 ...
##  $ Fouls.made               : num [1:20] 385 420 503 534 502 528 555 552 465 490 ...
##  $ Matches.without.conceding: num [1:20] 18 14 24 17 13 10 11 12 13 12 ...
##  $ Yellow.cards             : num [1:20] 66 72 91 100 84 116 106 110 108 110 ...
##  $ Red.cards                : num [1:20] 1 5 3 4 5 6 7 5 5 3 ...
##  $ Offsides                 : num [1:20] 120 114 84 106 92 103 106 85 85 80 ...
##  $ Effectiveness.Percentage : num [1:20] 64.8 57.4 51.8 51.7 52.4 ...
##  $ Team.Level               : chr [1:20] "Top" "Top" "Top" "Top" ...
```

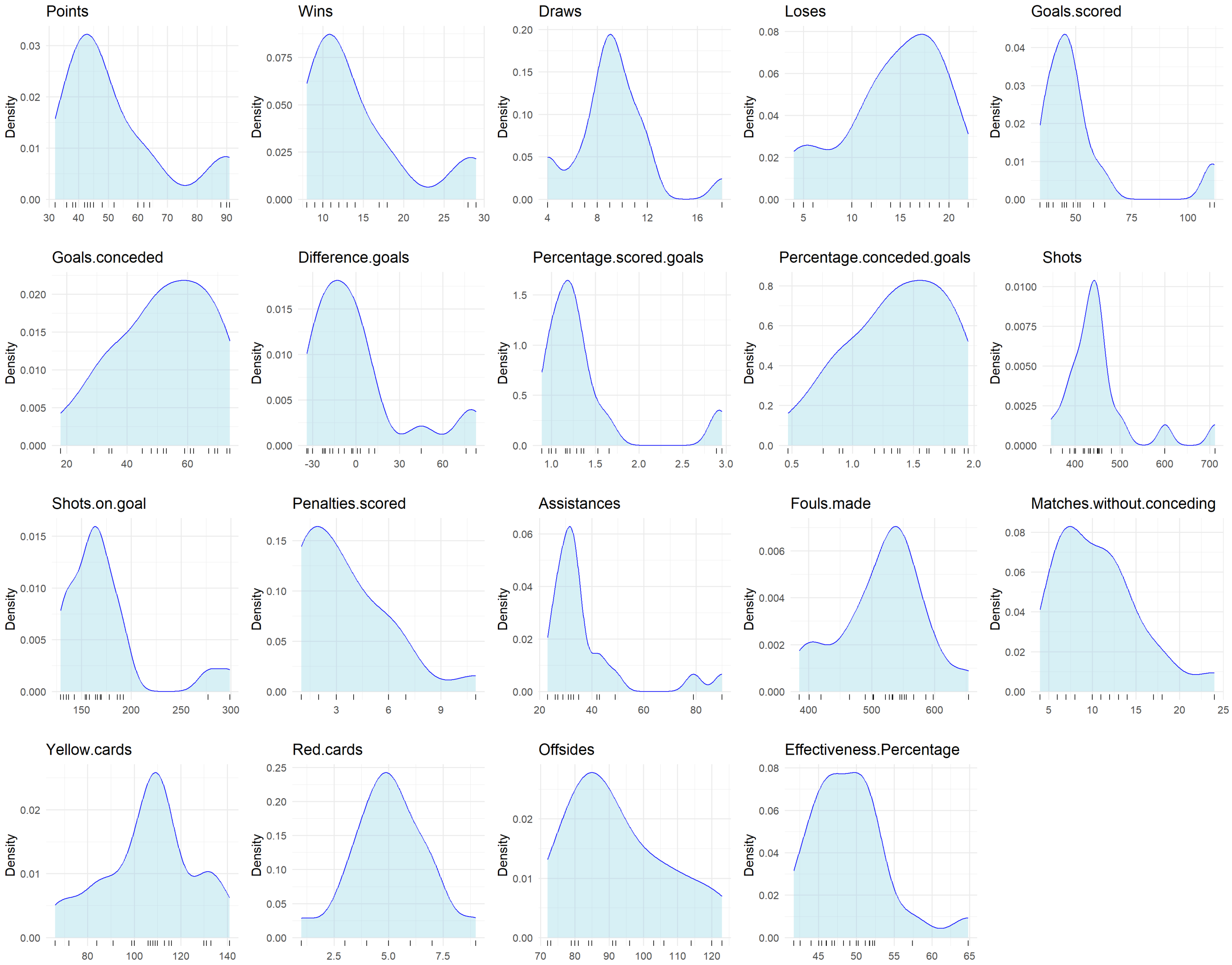
```
data
```

```
## # A tibble: 20 × 22
##    Team            Points Matches  Wins Draws Loses Goals.scored Goals.conceded
##    <chr>            <dbl>   <dbl> <dbl> <dbl> <dbl>        <dbl>          <dbl>
##  1 Barcelona           91      38    29     4     5          112             29
##  2 Real Madrid         90      38    28     6     4          110             34
##  3 Atlético Madrid     88      38    28     4     6           63             18
##  4 Villarreal          64      38    18    10    10           44             35
##  5 Athletic            62      38    18     8    12           58             45
##  6 Celta               60      38    17     9    12           51             59
##  7 Sevilla             52      38    14    10    14           51             50
##  8 Málaga              48      38    12    12    14           38             35
##  9 Real Sociedad       48      38    13     9    16           45             48
## 10 Betis               45      38    11    12    15           34             52
## 11 Las Palmas          44      38    12     8    18           45             53
## 12 Valencia            44      38    11    11    16           46             48
## 13 Eibar               43      38    11    10    17           49             61
## 14 Espanyol            43      38    12     7    19           40             74
## 15 Deportivo           42      38     8    18    12           45             61
## 16 Granada             39      38    10     9    19           46             69
## 17 Sporting Gijón      39      38    10     9    19           40             62
## 18 Rayo Vallecano      38      38     9    11    18           52             73
## 19 Getafe              36      38     9     9    20           37             67
## 20 Levante             32      38     8     8    22           37             70
## # ℹ 14 more variables: Difference.goals <dbl>, Percentage.scored.goals <dbl>,
## #   Percentage.conceded.goals <dbl>, Shots <dbl>, Shots.on.goal <dbl>,
## #   Penalties.scored <dbl>, Assistances <dbl>, Fouls.made <dbl>,
## #   Matches.without.conceding <dbl>, Yellow.cards <dbl>, Red.cards <dbl>,
## #   Offsides <dbl>, Effectiveness.Percentage <dbl>, Team.Level <chr>
```

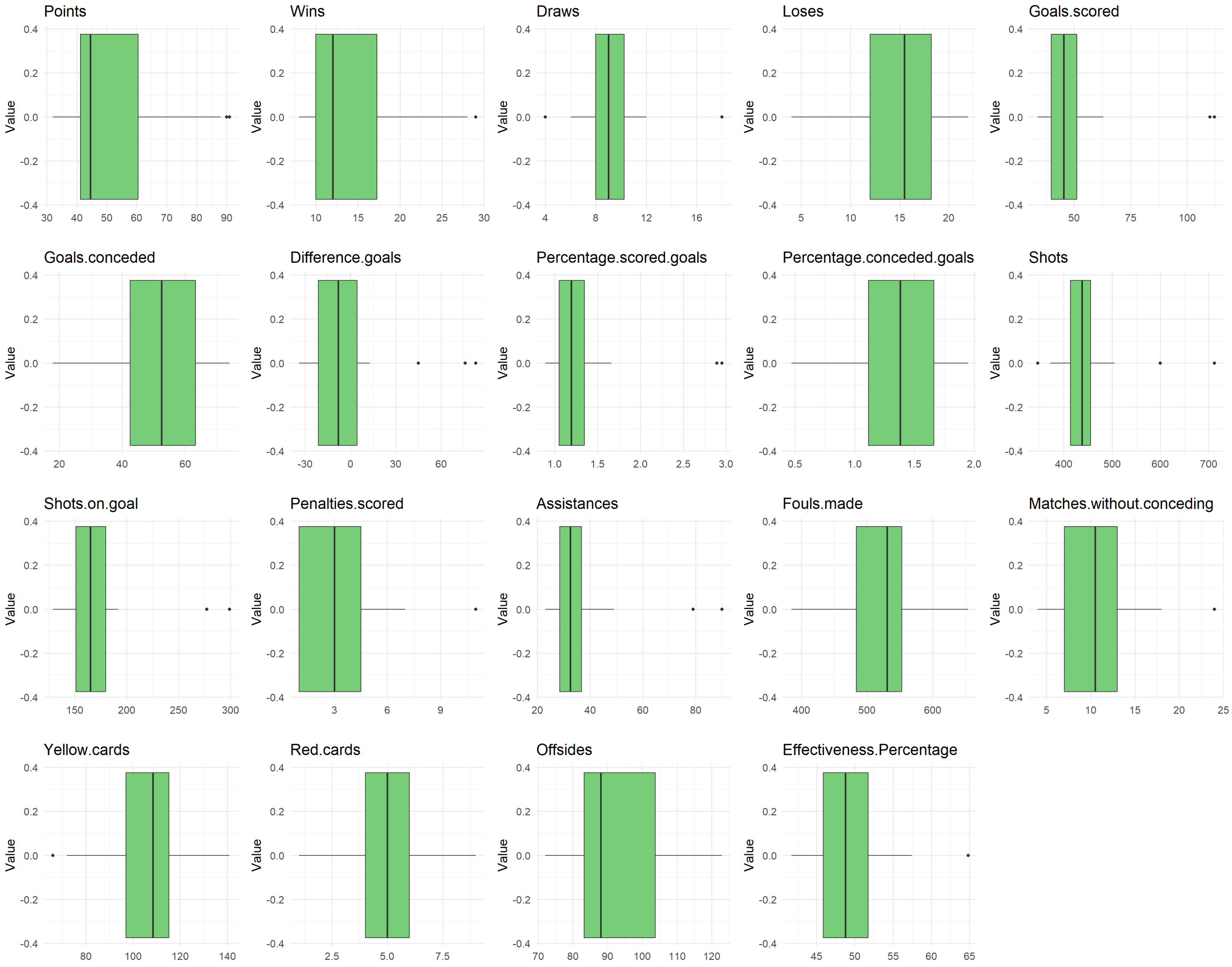
- Plot the distribution of numerical columns

```
plot_all_densities <- function(data, exclude_cols = c(1,3,22) , fill_color = "skyblue", alpha_value = 0.3, base_size = 15) {
  plot_list <- list()
  for (var in names(data[, -exclude_cols])) {
    p = ggplot(data, aes_string(x = var)) +
      geom_density(fill = "skyblue", color = "blue", alpha = alpha_value) +
      theme_minimal(base_size = base_size) +
      labs(title = var,
           x = "",
           y = "Density",
           ) +
      geom_rug(sides = "b")
     plot_list[[var]] <- p
  }
  title_grob <- textGrob("Distribution Of Numerical Columns", gp = gpar(fontsize = 20, fontface = "bold"))
  grid.arrange(title_grob, grobs=plot_list, ncol = 5, nrow = 5)
}

plot_all_densities(data)
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## ℹ Please use tidy evaluation idioms with `aes()`.
## ℹ See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```


Points
Wins
Draws
Loses
Goals.scored
Goals.conceded
Difference.goals
Percentage.scored.goals
Percentage.conceded.goals
Shots
Shots.on.goal
Penalties.scored
Assistances
Fouls.made
Matches.without.conceding
Yellow.cards
Red.cards
Offsides
Effectiveness.Percentage
Density
0.00
0.01
0.02
0.03
30 40 50 60 70 80 90
0.000
0.025
0.050
0.075
10 15 20 25 30
0.05
0.10
0.15
0.20
4 8 12 16
0.04
0.06
0.08
5 10 15 20
0.04
50 75 100
0.005
0.010
0.015
0.020
20 40 60
-30 0 30 60
0.0
0.5
1.0
1.5
1.0 1.5 2.0 2.5 3.0
0.2
0.4
0.6
0.8
0.5 1.0 1.5 2.0
0.0000
0.0025
0.0050
0.0075
0.0100
400 500 600 700
150 200 250 300
3 6 9
20 40 60 80
0.002
0.004
0.006
400 500 600
5 10 15 20 25
80 100 120 140
2.5 5.0 7.5
0.25
70 80 90 100 110 120
45 50 55 60 65

```
plot_all_boxplot <- function(data, exclude_cols = c(1, 3, 22), fill_color = "skyblue", alpha_value = 0.3, base_size = 15) {
  plot_list <- list()
  for (var in names(data[, -exclude_cols])) {
    p = ggplot(data, aes_string(x = var)) +
      geom_boxplot(fill='palegreen3') +
      theme_minimal(base_size = base_size) +
      labs(title = var,
           x = "",
           y = "Value"
           )
    plot_list[[var]] <- p
  }
  grid.arrange(title_grob,
               grobs=plot_list,
               ncol = 5,
               nrow = 5)
}

plot_all_boxplot(data)
```



These plots provide an overview of the distribution of football statistics in La Liga, showing a fairly even competition between teams, with some teams standing out. Points, wins, losses, and goals scored are all clustered around the median, with some teams like Barcelona, Real Madrid, or Atlético Madrid potentially having significantly higher metrics. Draws and fouls made are also tightly distributed, with a few teams tending to draw more or commit more fouls. Goal difference and win ratio percentage highlight a clear distinction between stronger and weaker teams, with some teams having notably high goal differences and win ratios. Goals conceded, offsides, yellow cards, and red cards also show a relatively tight distribution, reflecting the disciplined and strategic play of the teams. Metrics such as shots, shots on goal, and assists are evenly distributed, with some teams having notably more effective offenses. Overall, we illustrate the balance and intense competition in La Liga, with a few teams excelling in many key statistics.

2.**Comparison between the team levels** ()

```
top_teams <- data %>% filter(Team.Level == 'Top')
middle_teams <- data %>% filter(Team.Level == 'Middle')
bottom_teams <- data %>% filter(Team.Level == 'Bottom')
summary(top_teams)
```

```
##      Team               Points         Matches        Wins        Draws
##  Length:6           Min.   :60.00   Min.   :38   Min.   :17   Min.   : 4.000
##  Class :character   1st Qu.:62.50   1st Qu.:38   1st Qu.:18   1st Qu.: 4.500
##  Mode  :character   Median :76.00   Median :38   Median :23   Median : 7.000
##                     Mean   :75.83   Mean   :38   Mean   :23   Mean   : 6.833
##                     3rd Qu.:89.50   3rd Qu.:38   3rd Qu.:28   3rd Qu.: 8.750
##                     Max.   :91.00   Max.   :38   Max.   :29   Max.   :10.000
##      Loses        Goals.scored    Goals.conceded  Difference.goals
##  Min.   : 4.000   Min.   : 44.00   Min.   :18.00   Min.   :-8.00
##  1st Qu.: 5.250   1st Qu.: 52.75   1st Qu.:30.25   1st Qu.:10.00
##  Median : 8.000   Median : 60.50   Median :34.50   Median :29.00
##  Mean   : 8.167   Mean   : 73.00   Mean   :36.67   Mean   :36.33
##  3rd Qu.:11.500   3rd Qu.: 98.25   3rd Qu.:42.50   3rd Qu.:68.25
##  Max.   :12.000   Max.   :112.00   Max.   :59.00   Max.   :83.00
##  Percentage.scored.goals Percentage.conceded.goals     Shots
##  Min.   :1.160           Min.   :0.4700            Min.   :346.0
##  1st Qu.:1.387           1st Qu.:0.7925            1st Qu.:444.0
##  Median :1.595           Median :0.9050            Median :465.5
##  Mean   :1.922           Mean   :0.9617            Mean   :505.2
##  3rd Qu.:2.583           3rd Qu.:1.1150            3rd Qu.:570.2
##  Max.   :2.950           Max.   :1.5500            Max.   :712.0
##  Shots.on.goal   Penalties.scored  Assistances      Fouls.made
##  Min.   :135.0   Min.   : 1.000   Min.   :32.00   Min.   :385.0
##  1st Qu.:172.0   1st Qu.: 3.000   1st Qu.:42.25   1st Qu.:440.5
##  Median :182.0   Median : 3.500   Median :46.00   Median :502.5
##  Mean   :207.5   Mean   : 4.667   Mean   :55.83   Mean   :478.7
##  3rd Qu.:254.2   3rd Qu.: 5.500   3rd Qu.:71.50   3rd Qu.:521.8
##  Max.   :299.0   Max.   :11.000   Max.   :90.00   Max.   :534.0
##  Matches.without.conceding  Yellow.cards      Red.cards       Offsides
##  Min.   :10.00             Min.   : 66.00   Min.   :1.00   Min.   : 84.00
##  1st Qu.:13.25             1st Qu.: 75.00   1st Qu.:3.25   1st Qu.: 94.75
##  Median :15.50             Median : 87.50   Median :4.50   Median :104.50
##  Mean   :16.00             Mean   : 88.17   Mean   :4.00   Mean   :103.17
##  3rd Qu.:17.75             3rd Qu.: 97.75   3rd Qu.:5.00   3rd Qu.:112.00
##  Max.   :24.00             Max.   :116.00   Max.   :6.00   Max.   :120.00
##  Effectiveness.Percentage  Team.Level
##  Min.   :50.00            Length:6
##  1st Qu.:51.74            Class :character
##  Median :52.10            Mode  :character
##  Mean   :54.70
##  3rd Qu.:56.19
##  Max.   :64.83
```

```
cat('\n \n')
```

```
summary(middle_teams)
```

```
##      Team               Points         Matches        Wins          Draws      
##  Length:7           Min.   :43.00   Min.   :38   Min.   :11.0   Min.   : 8.00  
##  Class :character   1st Qu.:44.00   1st Qu.:38   1st Qu.:11.0   1st Qu.: 9.50  
##  Mode  :character   Median :45.00   Median :38   Median :12.0   Median :10.00  
##                     Mean   :46.29   Mean   :38   Mean   :12.0   Mean   :10.29  
##                     3rd Qu.:48.00   3rd Qu.:38   3rd Qu.:12.5   3rd Qu.:11.50  
##                     Max.   :52.00   Max.   :38   Max.   :14.0   Max.   :12.00  
##      Loses        Goals.scored  Goals.conceded  Difference.goals
##  Min.   :14.00   Min.   :34.0   Min.   :35.00   Min.   :-18.000 
##  1st Qu.:14.50   1st Qu.:41.5   1st Qu.:48.00   1st Qu.:-10.000 
##  Median :16.00   Median :45.0   Median :50.00   Median : -3.000 
##  Mean   :15.71   Mean   :44.0   Mean   :49.57   Mean   : -5.571 
##  3rd Qu.:16.50   3rd Qu.:47.5   3rd Qu.:52.50   3rd Qu.: -0.500 
##  Max.   :18.00   Max.   :51.0   Max.   :61.00   Max.   :  3.000 
##  Percentage.scored.goals Percentage.conceded.goals     Shots      
##  Min.   :0.890           Min.   :0.920             Min.   :398.0  
##  1st Qu.:1.090           1st Qu.:1.260             1st Qu.:409.5  
##  Median :1.180           Median :1.320             Median :421.0  
##  Mean   :1.156           Mean   :1.304             Mean   :429.1  
##  3rd Qu.:1.250           3rd Qu.:1.380             3rd Qu.:453.0  
##  Max.   :1.340           Max.   :1.610             Max.   :460.0  
##  Shots.on.goal   Penalties.scored  Assistances      Fouls.made   
##  Min.   :132.0   Min.   :1.000    Min.   :26.00   Min.   :465.0  
##  1st Qu.:150.5   1st Qu.:2.500    1st Qu.:27.00   1st Qu.:477.5  
##  Median :164.0   Median :4.000    Median :33.00   Median :552.0  
##  Mean   :160.7   Mean   :3.714    Mean   :30.57   Mean   :527.4  
##  3rd Qu.:169.5   3rd Qu.:5.000    3rd Qu.:33.00   3rd Qu.:561.0  
##  Max.   :189.0   Max.   :6.000    Max.   :35.00   Max.   :598.0  
##  Matches.without.conceding  Yellow.cards     Red.cards    Offsides     
##  Min.   : 7.00             Min.   : 99.0   Min.   :3   Min.   : 79.00  
##  1st Qu.: 9.00             1st Qu.:107.0   1st Qu.:4   1st Qu.: 82.50  
##  Median :11.00             Median :109.0   Median :5   Median : 85.00  
##  Mean   :10.43             Mean   :107.9   Mean   :5   Mean   : 87.43  
##  3rd Qu.:12.00             3rd Qu.:110.0   3rd Qu.:6   3rd Qu.: 88.50  
##  Max.   :13.00             Max.   :113.0   Max.   :7   Max.   :106.00  
##  Effectiveness.Percentage  Team.Level       
##  Min.   :41.71            Length:7          
##  1st Qu.:46.03            Class :character  
##  Median :47.13            Mode  :character  
##  Mean   :47.63                              
##  3rd Qu.:50.19                              
##  Max.   :52.17                              
```

```
cat('\n \n')
```

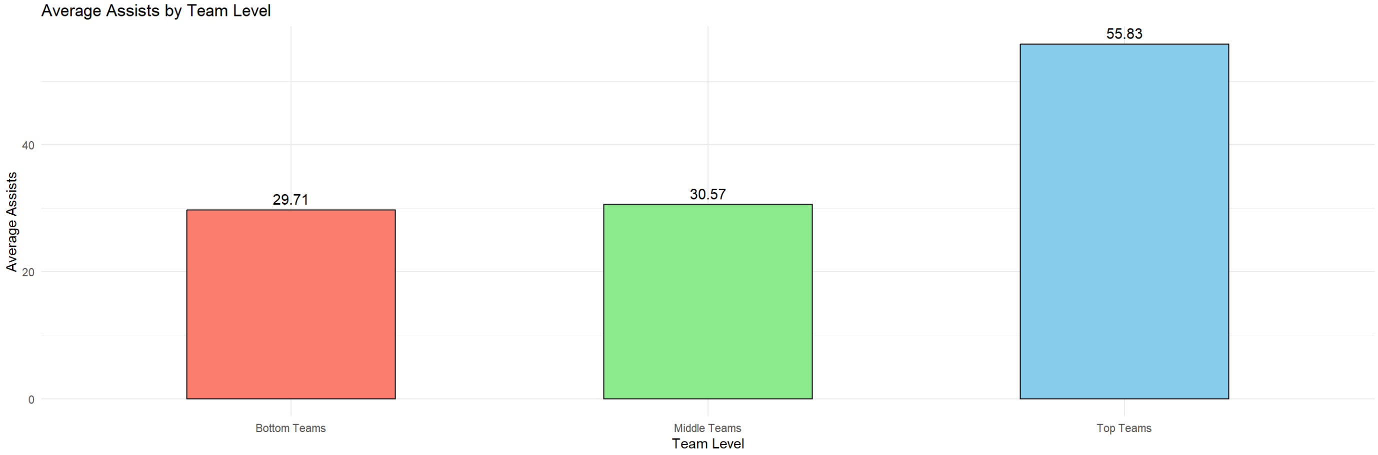
```
summary(bottom_teams)
```

```
##      Team               Points         Matches        Wins       
##  Length:7           Min.   :32.00   Min.   :38   Min.   : 8.000  
##  Class :character   1st Qu.:37.00   1st Qu.:38   1st Qu.: 8.500  
##  Mode  :character   Median :39.00   Median :38   Median : 9.000  
##                     Mean   :38.43   Mean   :38   Mean   : 9.429  
##                     3rd Qu.:40.50   3rd Qu.:38   3rd Qu.:10.000  
##                     Max.   :43.00   Max.   :38   Max.   :12.000  
##      Draws           Loses        Goals.scored   Goals.conceded
##  Min.   : 7.00   Min.   :12.00   Min.   :37.00   Min.   :61.0  
##  1st Qu.: 8.50   1st Qu.:18.50   1st Qu.:38.50   1st Qu.:64.5  
##  Median : 9.00   Median :19.00   Median :40.00   Median :69.0  
##  Mean   :10.14   Mean   :18.43   Mean   :42.43   Mean   :68.0  
##  3rd Qu.:10.00   3rd Qu.:19.50   3rd Qu.:45.50   3rd Qu.:71.5  
##  Max.   :18.00   Max.   :22.00   Max.   :52.00   Max.   :74.0  
##  Difference.goals Percentage.scored.goals Percentage.conceded.goals
##  Min.   :-34.00   Min.   :0.970           Min.   :1.610            
##  1st Qu.:-31.50   1st Qu.:1.010           1st Qu.:1.695            
##  Median :-23.00   Median :1.050           Median :1.820            
##  Mean   :-25.57   Mean   :1.114           Mean   :1.790            
##  3rd Qu.:-21.50   3rd Qu.:1.195           3rd Qu.:1.880            
##  Max.   :-16.00   Max.   :1.370           Max.   :1.950            
##      Shots       Shots.on.goal   Penalties.scored  Assistances   
##  Min.   :372.0   Min.   :129.0   Min.   :1.000    Min.   :23.00  
##  1st Qu.:409.0   1st Qu.:145.5   1st Qu.:1.000    1st Qu.:28.00  
##  Median :433.0   Median :155.0   Median :1.000    Median :31.00  
##  Mean   :430.4   Mean   :155.9   Mean   :2.143    Mean   :29.71  
##  3rd Qu.:442.5   3rd Qu.:162.0   3rd Qu.:2.000    3rd Qu.:31.50  
##  Max.   :505.0   Max.   :192.0   Max.   :7.000    Max.   :35.00  
##    Fouls.made    Matches.without.conceding  Yellow.cards     Red.cards  
##  Min.   :401.0   Min.   :4.000             Min.   : 84.0   Min.   :4.0  
##  1st Qu.:527.5   1st Qu.:6.000             1st Qu.:111.0   1st Qu.:5.0  
##  Median :545.0   Median :7.000             Median :130.0   Median :6.0  
##  Mean   :541.3   Mean   :6.429             Mean   :120.1   Mean   :6.0  
##  3rd Qu.:567.0   3rd Qu.:7.000             3rd Qu.:132.0   3rd Qu.:6.5  
##  Max.   :654.0   Max.   :8.000             Max.   :141.0   Max.   :9.0  
##     Offsides      Effectiveness.Percentage  Team.Level       
##  Min.   : 72.00   Min.   :42.56            Length:7          
##  1st Qu.: 77.00   1st Qu.:44.52            Class :character  
##  Median : 85.00   Median :45.43            Mode  :character  
##  Mean   : 88.71   Mean   :46.06                              
##  3rd Qu.: 93.50   3rd Qu.:47.55                              
##  Max.   :123.00   Max.   :50.26                              
```

2.1. Attack and defense statistics ()

- 2.1.1. Average asisstances by team level ()

```
avg_assists_top <- round(mean(top_teams$Assistances), digits = 2)
avg_assists_middle <- round(mean(middle_teams$Assistances), digits = 2)
avg_assists_bottom <- round(mean(bottom_teams$Assistances), digits = 2)

avg_assists <- data.frame(
  Team_Level = c("Top Teams", "Middle Teams", "Bottom Teams"),
  Avg_Assists = c(avg_assists_top, avg_assists_middle, avg_assists_bottom)
)
ggplot(avg_assists, aes(x = Team_Level, y = Avg_Assists, fill = Team_Level)) +
      geom_bar(stat = "identity", width = 0.5, color = "black") +
      geom_text(aes(label = Avg_Assists), vjust = -0.5, color = "black", size = 4) +
      labs(title = "Average Assists by Team Level",
           x = "Team Level",
           y = "Average Assists",
           fill = "Team Level") +
      theme_minimal() +
      theme(legend.position = "none") +
      scale_fill_manual(values = c("Top Teams" = "skyblue", "Middle Teams" = "lightgreen", "Bottom Teams" = "salmon"))
```



- 2.1.2. Comparison of Shots, Shots on Goal, and Goals Scored by Team Level. ()

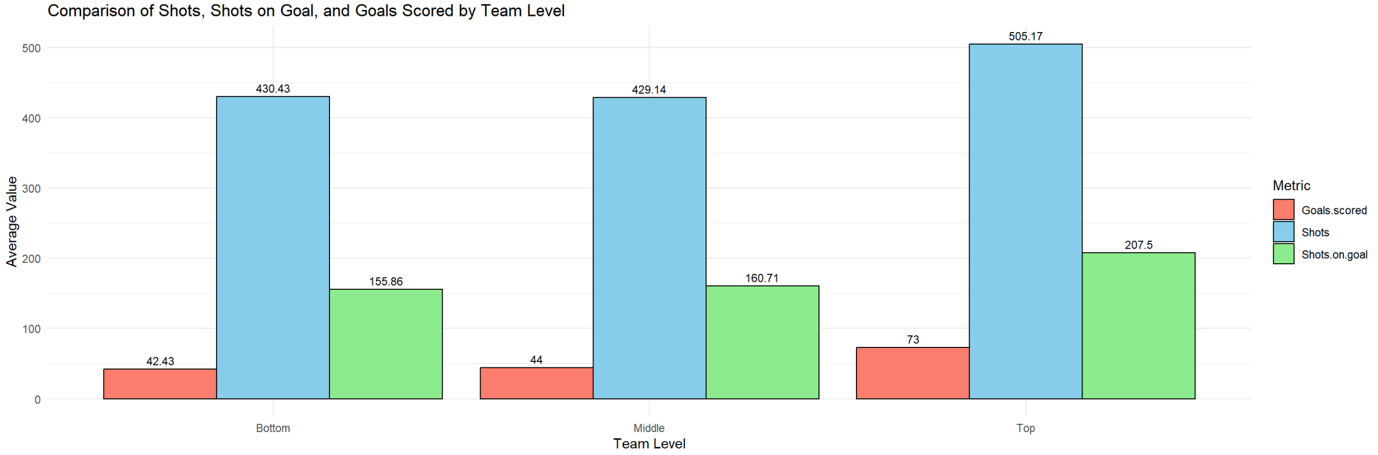
```
long_data <- data %>%
  gather(key = "Metric", value = "Value", Shots, Shots.on.goal, Goals.scored)
avg_metrics <- long_data %>%
  group_by(Team.Level, Metric) %>%
  summarise(Average = mean(Value))
```

```
## `summarise()` has grouped output by 'Team.Level'. You can override using the
## `.groups` argument.
```

```
avg_metrics
```

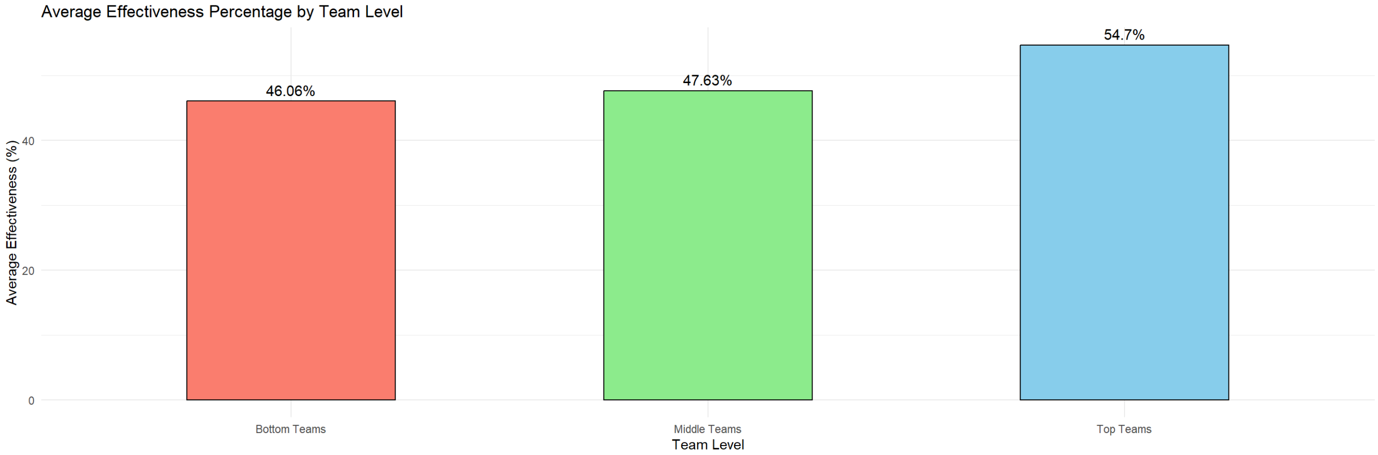
```
## # A tibble: 9 × 3
## # Groups:   Team.Level [3]
##   Team.Level Metric        Average
##   <chr>      <chr>           <dbl>
## 1 Bottom     Goals.scored     42.4
## 2 Bottom     Shots           430.
## 3 Bottom     Shots.on.goal   156.
## 4 Middle     Goals.scored     44
## 5 Middle     Shots           429.
## 6 Middle     Shots.on.goal   161.
## 7 Top        Goals.scored     73
## 8 Top        Shots           505.
## 9 Top        Shots.on.goal   208.
```

```
ggplot(avg_metrics, aes(x = Team.Level, y = Average, fill = Metric)) +
  geom_bar(stat = "identity", position = position_dodge(), color = "black") +
  geom_text(aes(label = round(Average, 2)), position = position_dodge(0.9), vjust = -0.5, size = 3) +
  labs(title = "Comparison of Shots, Shots on Goal, and Goals Scored by Team Level",
       x = "Team Level",
       y = "Average Value",
       fill = "Metric") +
  theme_minimal() +
  scale_fill_manual(values = c("Shots" = "skyblue", "Shots.on.goal" = "lightgreen", "Goals.scored" = "salmon"))
```



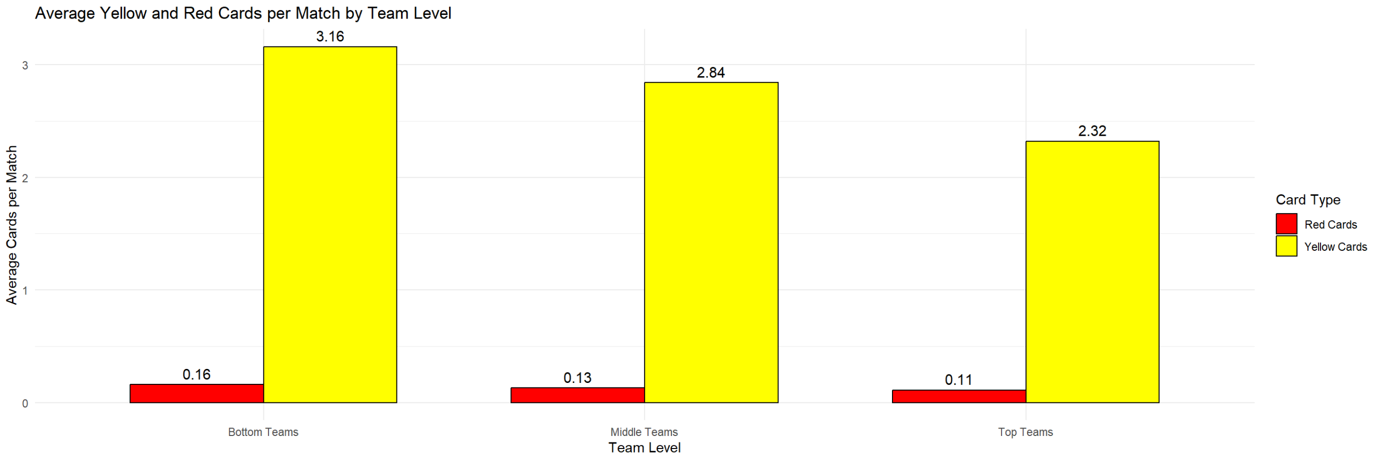
- 2.1.3. Do bottom teams tend to have less effective shots than top teams ? ()

```
avg_top <- round(mean(top_teams$Effectiveness.Percentage), digits=2)
avg_bottom <- round(mean(bottom_teams$Effectiveness.Percentage), digits=2)
avg_middle <- round(mean(middle_teams$Effectiveness.Percentage), digits=2)

avg_effectiveness <- data.frame(
  Team_Level = c("Top Teams", "Middle Teams" ,"Bottom Teams"),
  Avg_effectiveness = c(avg_top, avg_middle, avg_bottom)
)

ggplot(avg_effectiveness, aes(x = Team_Level, y = Avg_effectiveness, fill = Team_Level)) +
  geom_bar(stat = "identity", width = 0.5, color = "black") +
  geom_text(aes(label = paste0(Avg_effectiveness, "%")), vjust = -0.5, color = "black", size = 4) +
  labs(title = "Average Effectiveness Percentage by Team Level",
       x = "Team Level",
       y = "Average Effectiveness (%)",
       fill = "Team Level") +
  theme_minimal() +
  theme(legend.position = "none") +
  scale_fill_manual(values = c("Top Teams" = "skyblue", "Middle Teams" = "lightgreen", "Bottom Teams" = "salmon"))
```



2.2. Analysis of penalty cards. ()

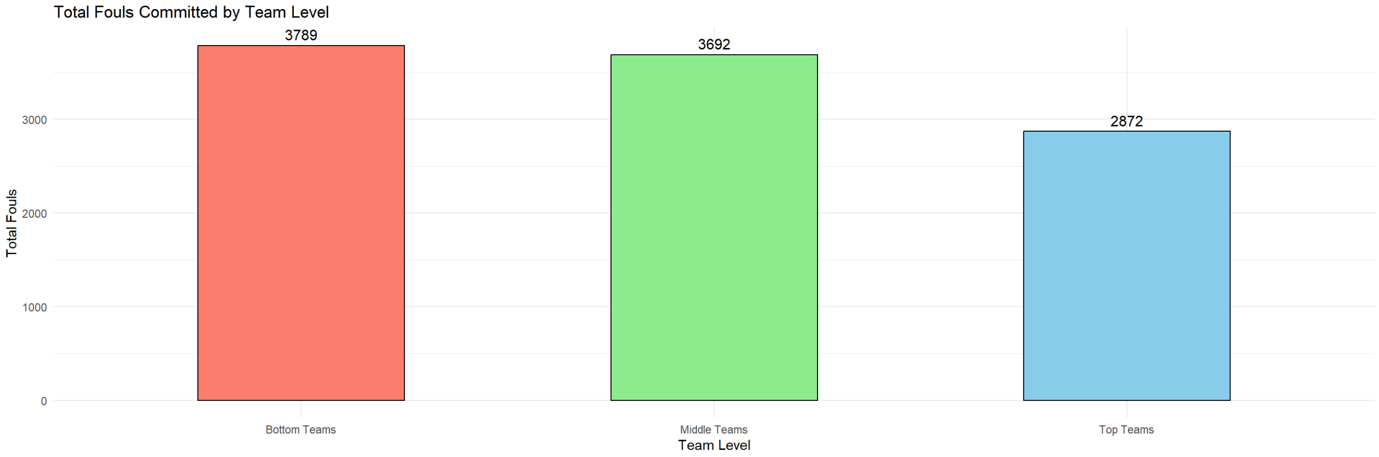
```
avg_yellow_cards_top <- round(mean(top_teams$Yellow.cards / top_teams$Matches), digits = 2)
avg_yellow_cards_middle <- round(mean(middle_teams$Yellow.cards / middle_teams$Matches), digits = 2)
avg_yellow_cards_bottom <- round(mean(bottom_teams$Yellow.cards / bottom_teams$Matches), digits = 2)

avg_red_cards_top <- round(mean(top_teams$Red.cards / top_teams$Matches), digits = 2)
avg_red_cards_middle <- round(mean(middle_teams$Red.cards / middle_teams$Matches), digits = 2)
avg_red_cards_bottom <- round(mean(bottom_teams$Red.cards / bottom_teams$Matches), digits = 2)

avg_cards <- data.frame(
  Team_Level = rep(c("Top Teams", "Middle Teams", "Bottom Teams"), each = 2),
  Card_Type = rep(c("Yellow Cards", "Red Cards"), times = 3),
  Avg_Cards = c(avg_yellow_cards_top, avg_red_cards_top,
                avg_yellow_cards_middle, avg_red_cards_middle,
                avg_yellow_cards_bottom, avg_red_cards_bottom)
)

ggplot(avg_cards, aes(x = Team_Level, y = Avg_Cards, fill = Card_Type)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7, color = "black") +
  geom_text(aes(label = Avg_Cards), position = position_dodge(width = 0.7), vjust = -0.5, size = 4) +
  labs(title = "Average Yellow and Red Cards per Match by Team Level",
       x = "Team Level",
       y = "Average Cards per Match",
       fill = "Card Type") +
  theme_minimal() +
  scale_fill_manual(values = c("Yellow Cards" = "yellow", "Red Cards" = "red"))
```



2.3. Analysis of tactical indicators: ()

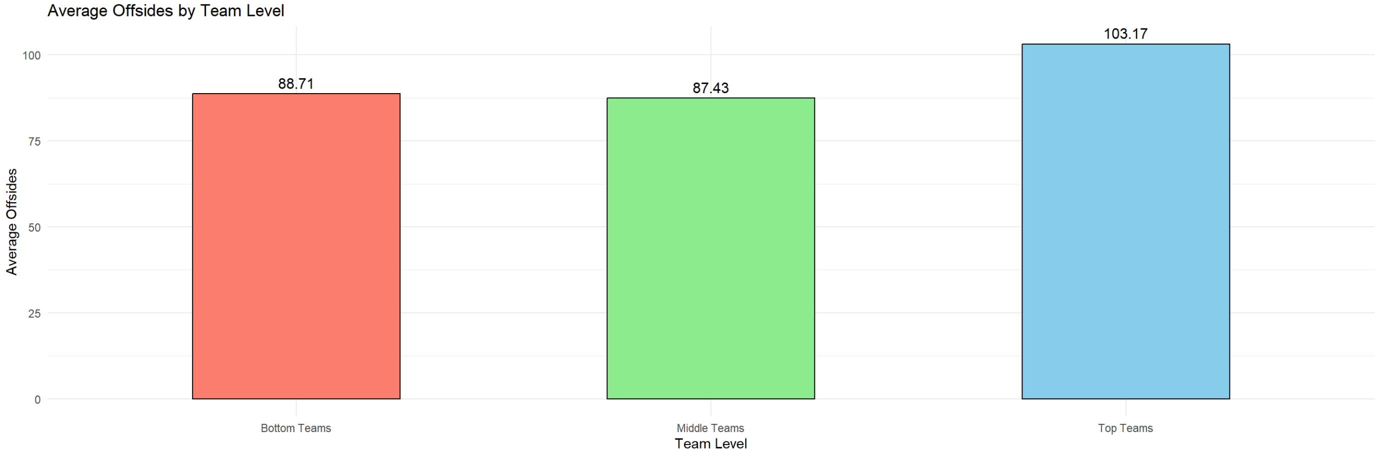
- 2.3.1. Do lower ranked teams commit more fouls than higher ranked teams? ()

```
total_fouls_top <- sum(top_teams$Fouls.made)
total_fouls_bottom <- sum(bottom_teams$Fouls.made)
total_fouls_middle <- sum(middle_teams$Fouls.made)

# Create a data frame for total fouls
total_fouls <- data.frame(
  Team_Level = c("Top Teams", "Middle Teams" ,"Bottom Teams"),
  Total_Fouls = c(total_fouls_top, total_fouls_middle ,total_fouls_bottom)
)

# Plot the bar chart
ggplot(total_fouls, aes(x = Team_Level, y = Total_Fouls, fill = Team_Level)) +
  geom_bar(stat = "identity", width = 0.5, color = "black") +
  geom_text(aes(label = Total_Fouls), vjust = -0.5, color = "black", size = 4) +
  labs(title = "Total Fouls Committed by Team Level",
       x = "Team Level",
       y = "Total Fouls",
       fill = "Team Level") +
  theme_minimal() +
  theme(legend.position = "none") +
  scale_fill_manual(values = c("Top Teams" = "skyblue", "Middle Teams" = "lightgreen", "Bottom Teams" = "salmon"))
```

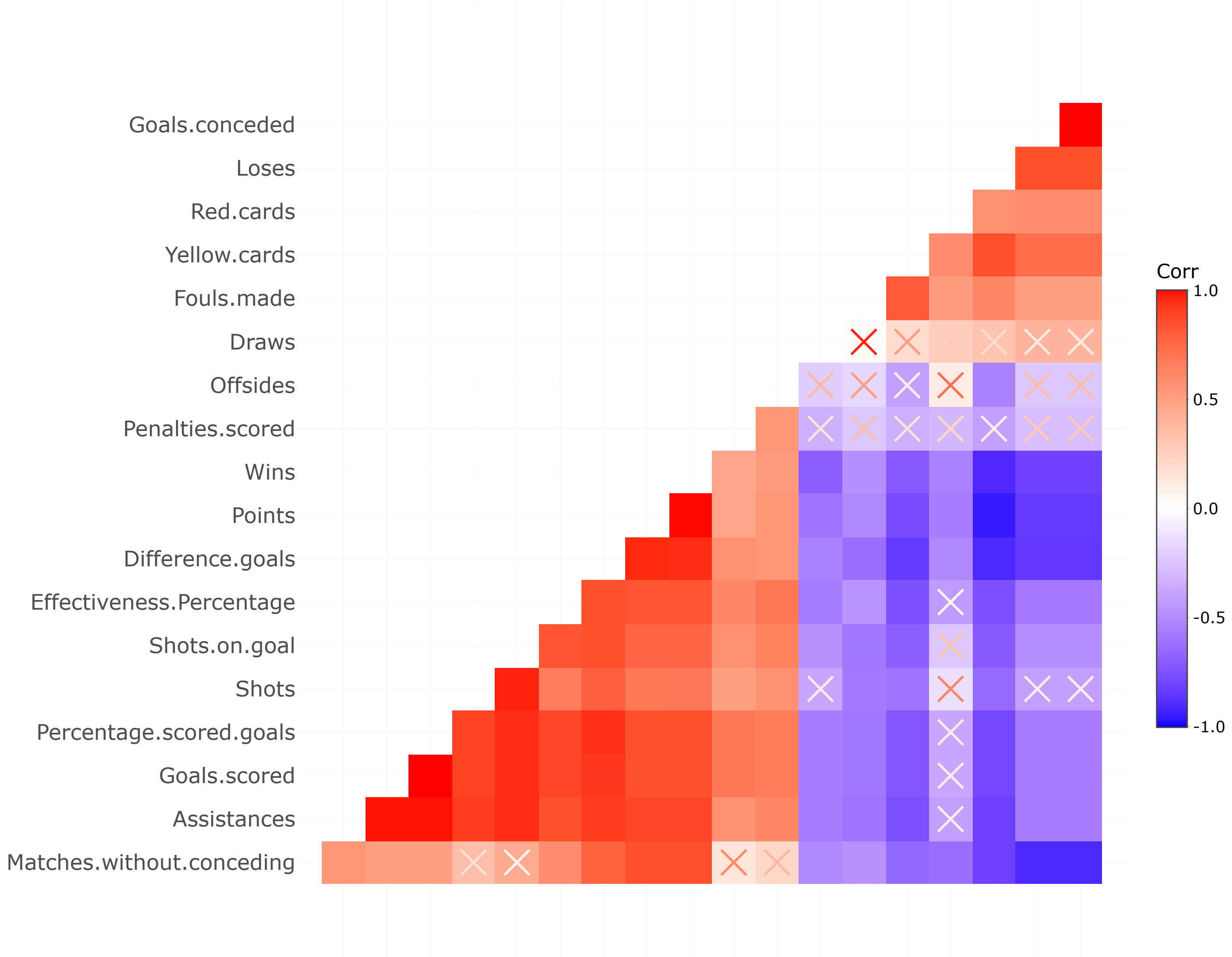


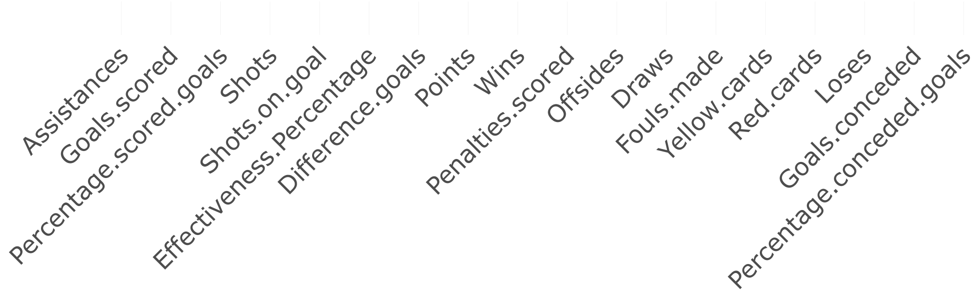
- 2.3.3. Average number of offsides committed that season ()

```
avg_offsides_top <- round(mean(top_teams$Offsides), digits = 2)
avg_offsides_middle <- round(mean(middle_teams$Offsides), digits = 2)
avg_offsides_bottom <- round(mean(bottom_teams$Offsides), digits = 2)

avg_offsides <- data.frame(
  Team_Level = c("Top Teams", "Middle Teams", "Bottom Teams"),
  Avg_Offsides = c(avg_offsides_top, avg_offsides_middle, avg_offsides_bottom)
)

ggplot(avg_offsides, aes(x = Team_Level, y = Avg_Offsides, fill = Team_Level)) +
  geom_bar(stat = "identity", width = 0.5, color = "black") +
  geom_text(aes(label = Avg_Offsides), vjust = -0.5, color = "black", size = 4) +
  labs(title = "Average Offsides by Team Level",
       x = "Team Level",
       y = "Average Offsides",
       fill = "Team Level") +
  theme_minimal() +
  theme(legend.position = "none") +
  scale_fill_manual(values = c("Top Teams" = "skyblue", "Middle Teams" = "lightgreen", "Bottom Teams" = "salmon"))
```



```
corr_mat <- round(cor(data[, -c(1,3,22)]),2)
p_mat <- cor_pmat(data[, -c(1,3,22)])

# plotting the interactive corr heatmap
corr_mat <- ggcorrplot(
  corr_mat, hc.order = TRUE, type = "lower",
  outline.col = "white",
  p.mat = p_mat
)

ggplotly(corr_mat)
```





3. Principal Components Analysis ()

```
data.pca <- PCA(data[, -c(1,3,22)], graph=F)
data.pca
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 20 individuals, described by 19 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"
```

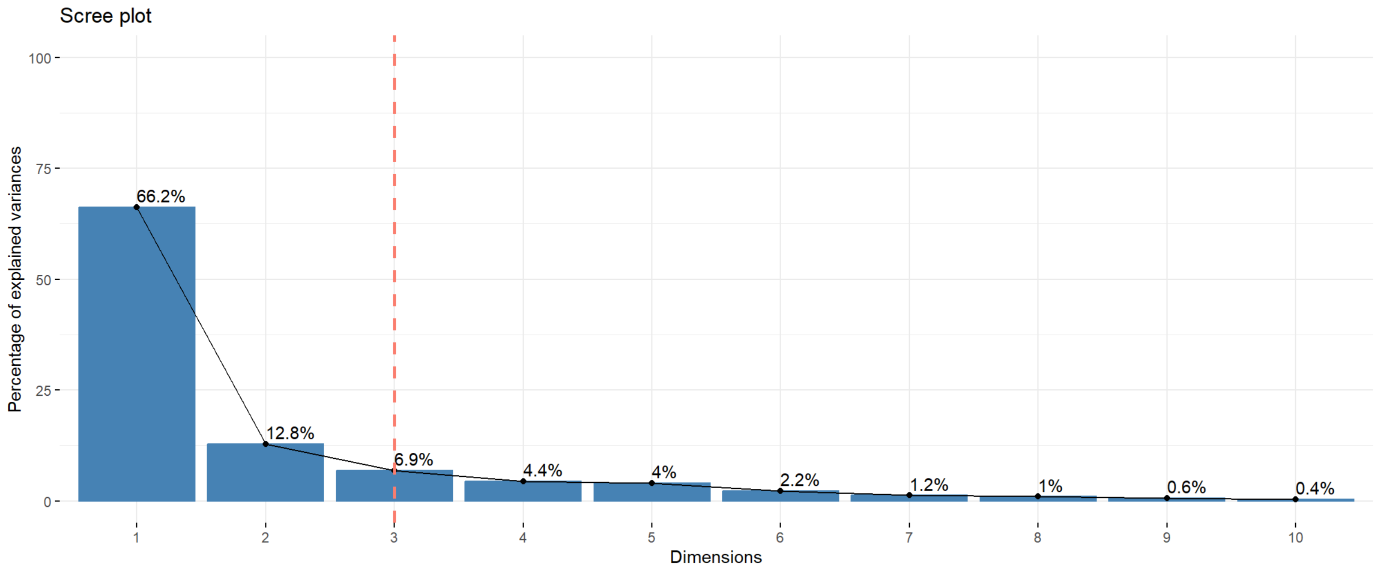
```
eig.val <- get_eigenvalue(data.pca)
eig.val
```

```
##          eigenvalue variance.percent cumulative.variance.percent
## Dim.1  1.257613e+01     6.619016e+01                    66.19016
## Dim.2  2.440067e+00     1.284246e+01                    79.03261
## Dim.3  1.310948e+00     6.899725e+00                    85.93234
## Dim.4  8.385734e-01     4.413544e+00                    90.34588
## Dim.5  7.582783e-01     3.990938e+00                    94.33682
## Dim.6  4.199656e-01     2.210345e+00                    96.54717
## Dim.7  2.354754e-01     1.239344e+00                    97.78651
## Dim.8  1.835014e-01     9.657967e-01                    98.75231
## Dim.9  1.114548e-01     5.866043e-01                    99.33891
## Dim.10 6.889043e-02     3.625812e-01                    99.70149
## Dim.11 3.523408e-02     1.854425e-01                    99.88693
## Dim.12 1.378479e-02     7.255153e-02                    99.95949
## Dim.13 7.014233e-03     3.691702e-02                    99.99640
## Dim.14 6.742830e-04     3.548858e-03                    99.99995
## Dim.15 7.732127e-06     4.069540e-05                    99.99999
## Dim.16 1.479589e-06     7.787311e-06                   100.00000
## Dim.17 2.494574e-31     1.312933e-30                   100.00000
## Dim.18 9.583700e-33     5.044052e-32                   100.00000
## Dim.19 7.740154e-33     4.073765e-32                   100.00000
```

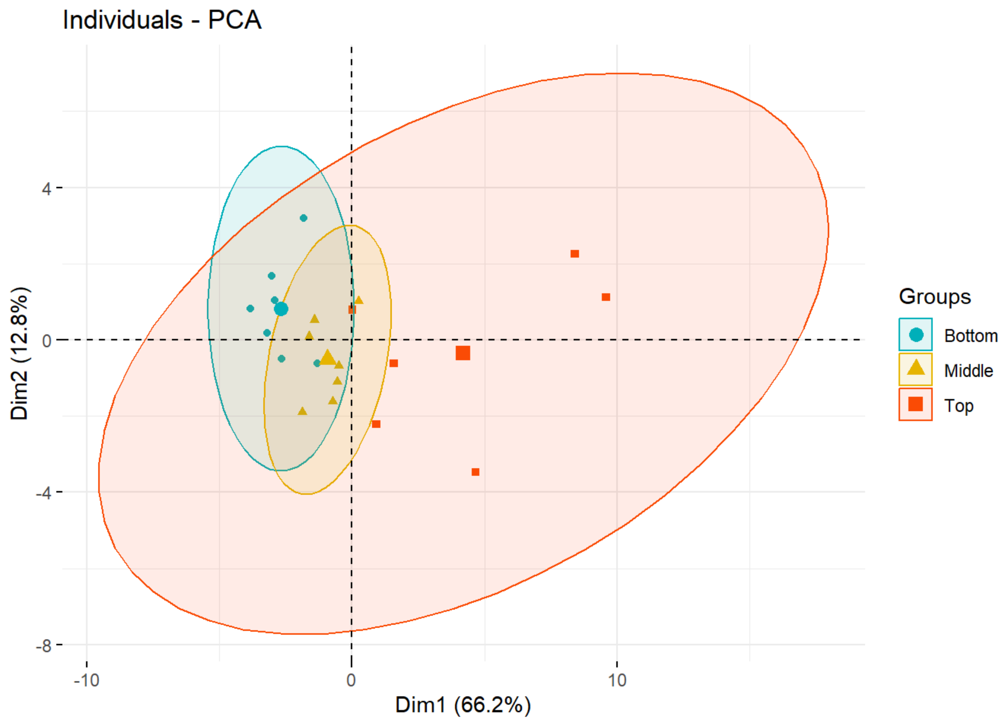
- The proportion of variance is represented by an eigenvalue in the second column, For example, Dim.1 has an eigenvalue of 15.393, which corresponds to a variance percentage of 66.92852
- We can limit the number of principal components to a certain fraction of the total variance (eg > 70%).

```
fviz_eig(data.pca, addlabels = TRUE, ylim = c(0, 100)) +
  geom_vline(xintercept = 3, linetype = "dashed", color = "salmon", size = 1)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```
fviz_pca_ind(data.pca,
             geom.ind = "point", # show points only (nbut not "text")
             col.ind = data$Team.Level, # color by groups
             palette = c("#00AFBB", "#E7B800", "#FC4E07"),
             addEllipses = TRUE, # Concentration ellipses
             legend.title = "Groups"
             )
```

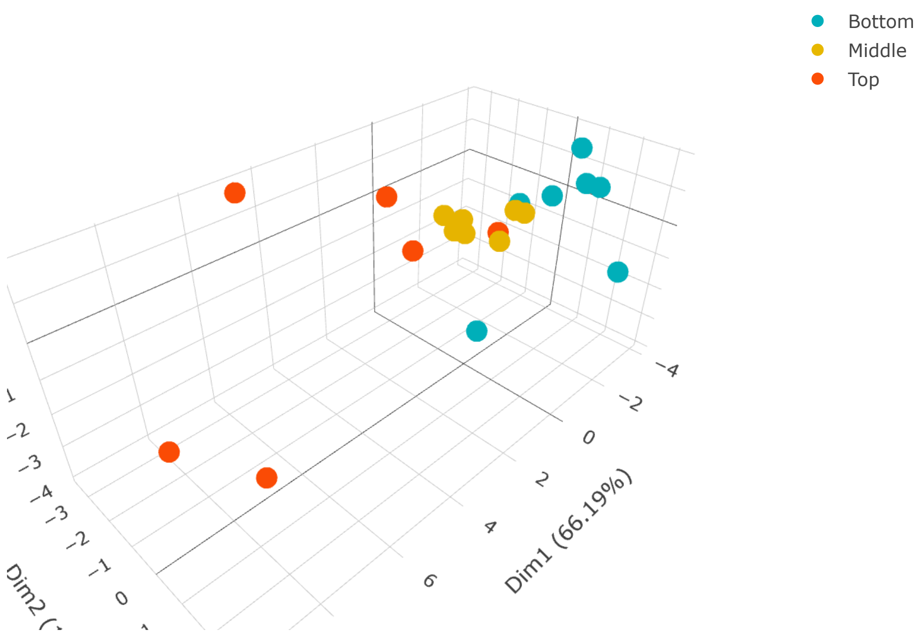


```
data.pca
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 20 individuals, described by 19 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"
```

```
explained_variance <- data.pca$eig[, 2]
pca_3d <- as.data.frame(data.pca$ind$coord)
pca_3d$Team.Level <- data$Team.Level

plot_ly(pca_3d, x = ~Dim.1, y = ~Dim.2, z = ~Dim.3, color = ~Team.Level, colors = c("#00AFBB", "#E7B800", "#FC4E07")) %>%
  add_markers() %>%
  layout(scene = list(xaxis = list(title = paste0('Dim1 (', round(explained_variance[1], 2), '%)')),
                      yaxis = list(title = paste0('Dim2 (', round(explained_variance[2], 2), '%)')),
                      zaxis = list(title = paste0('Dim3 (', round(explained_variance[3], 2), '%)'))),
         title = "3D PCA Plot of La Liga Teams")
```

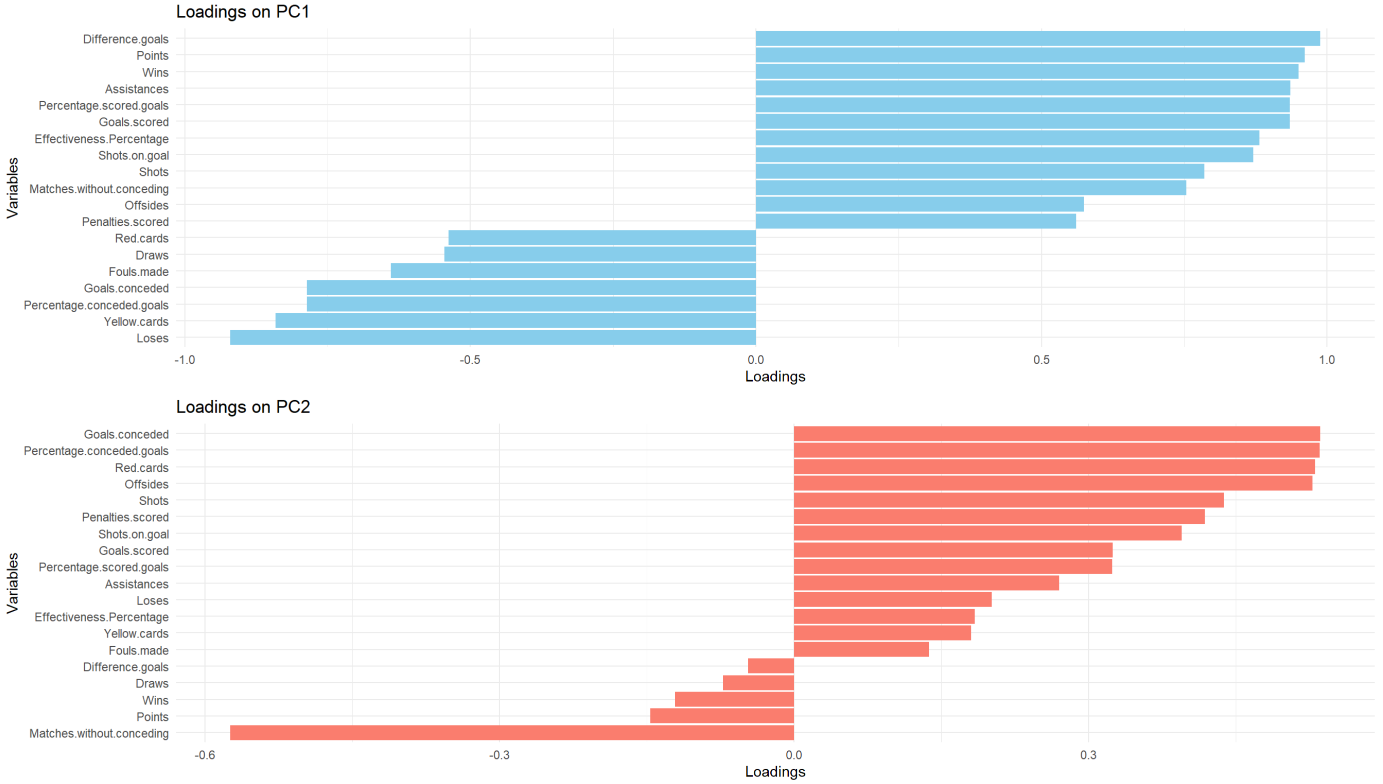
3D PCA Plot of La Liga Teams



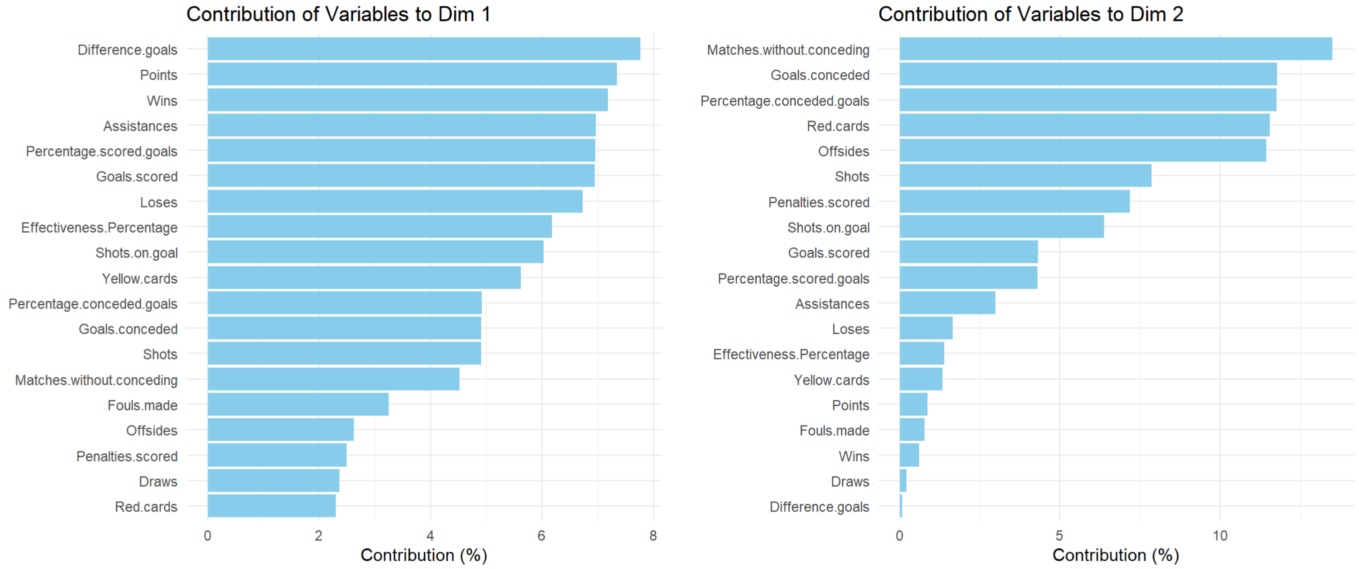
```
loadings <- as.data.frame(data.pca$var$coord)
loadings$Variable <- rownames(loadings)
loadings
```

```
##                                  Dim.1       Dim.2       Dim.3       Dim.4
## Points                     0.9610355 -0.14612550  0.14831165  0.08488793
## Wins                       0.9502959 -0.12095188  0.24193789  0.04219864
## Draws                     -0.5456861 -0.07247303 -0.71339332  0.22742322
## Loses                     -0.9199389  0.20154147  0.10762572 -0.19059739
## Goals.scored               0.9347944  0.32484416 -0.02716391 -0.08463933
## Goals.conceded            -0.7856982  0.53608151 -0.06893795 -0.16196976
## Difference.goals           0.9881537 -0.04676619  0.01553648  0.02293414
## Percentage.scored.goals    0.9352299  0.32401622 -0.02529418 -0.08351271
## Percentage.conceded.goals -0.7862606  0.53541203 -0.06984591 -0.16239848
## Shots                      0.7853454  0.43773449 -0.18135404  0.07079497
## Shots.on.goal              0.8708401  0.39473982 -0.11076218  0.03167907
## Penalties.scored           0.5606864  0.41870702  0.10741196 -0.39974717
## Assistances                0.9358765  0.27016711 -0.03252452 -0.04133849
## Fouls.made                -0.6392594  0.13758682  0.64679785  0.15528033
## Matches.without.conceding  0.7539423 -0.57383667  0.18911310  0.12899563
## Yellow.cards              -0.8406716  0.18027712  0.40229603 -0.01469528
## Red.cards                 -0.5381657  0.53096898  0.09592143  0.55862442
## Offsides                   0.5739901  0.52830822  0.06145683  0.38851630
## Effectiveness.Percentage   0.8812976  0.18431942  0.10940898 -0.05263299
##                                    Dim.5                  Variable
## Points                    -0.004011597                    Points
## Wins                      -0.051608265                      Wins
## Draws                      0.316865928                     Draws
## Loses                     -0.120956689                     Loses
## Goals.scored              -0.047228868              Goals.scored
## Goals.conceded            -0.072294248            Goals.conceded
## Difference.goals           0.004076686          Difference.goals
## Percentage.scored.goals   -0.045045421   Percentage.scored.goals
## Percentage.conceded.goals -0.068074579 Percentage.conceded.goals
## Shots                     -0.310997074                     Shots
## Shots.on.goal             -0.186965768             Shots.on.goal
## Penalties.scored           0.476279539          Penalties.scored
## Assistances               -0.152332976               Assistances
## Fouls.made                 0.180048091                Fouls.made
## Matches.without.conceding -0.040287580 Matches.without.conceding
## Yellow.cards              -0.076464634              Yellow.cards
## Red.cards                 -0.143688399                 Red.cards
## Offsides                   0.396105174                  Offsides
## Effectiveness.Percentage   0.165140994  Effectiveness.Percentage
```

```
p1 <- ggplot(loadings, aes(x = reorder(Variable, Dim.1), y = Dim.1)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_flip() +
  labs(title = "Loadings on PC1", x = "Variables", y = "Loadings") +
  theme_minimal()

p2 <- ggplot(loadings, aes(x = reorder(Variable, Dim.2), y = Dim.2)) +
  geom_bar(stat = "identity", fill = "salmon") +
  coord_flip() +
  labs(title = "Loadings on PC2", x = "Variables", y = "Loadings") +
  theme_minimal()

grid.arrange(p1, p2, nrow = 2)
```

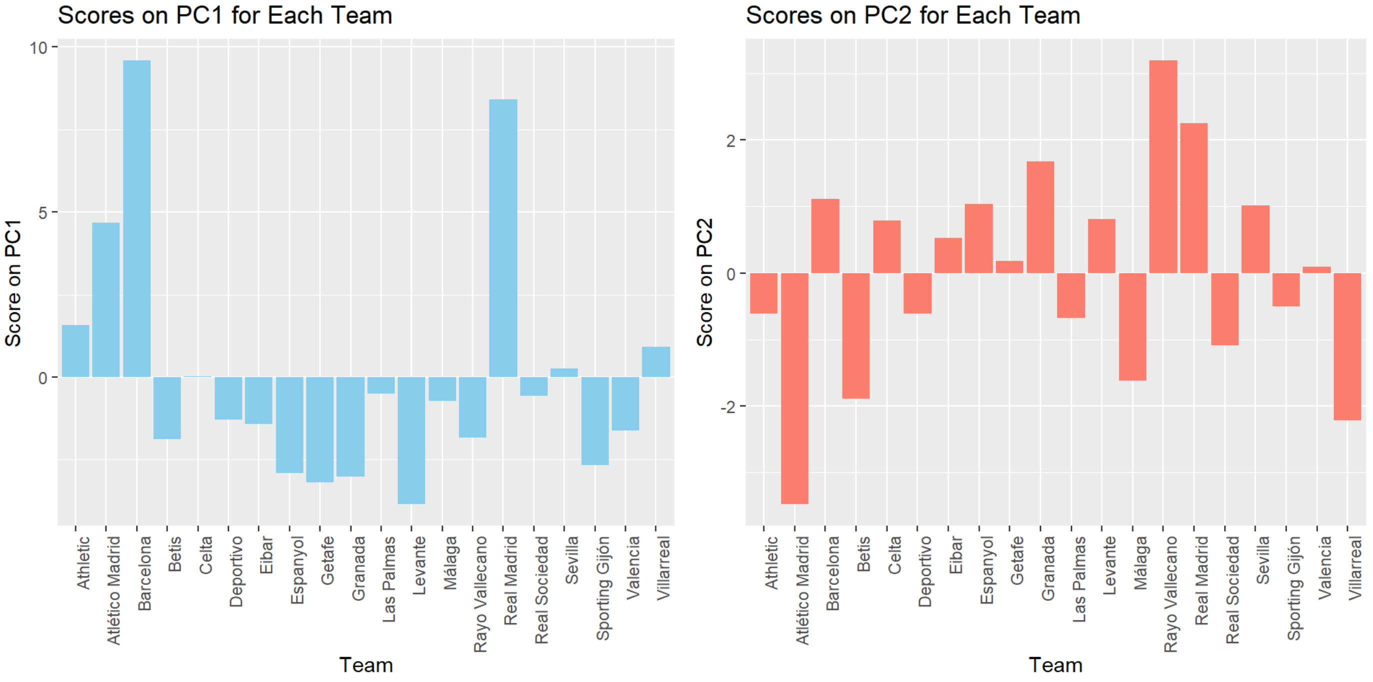


```
contrib_PC1 <- as.data.frame(data.pca$var$contrib[,1])
colnames(contrib_PC1) <- c("Contribution")
contrib_PC1$Variable <- rownames(contrib_PC1)

contrib_PC2 <- as.data.frame(data.pca$var$contrib[,2])
colnames(contrib_PC2) <- c("Contribution")
contrib_PC2$Variable <- rownames(contrib_PC2)

p1 = ggplot(contrib_PC1, aes(x = reorder(Variable, Contribution), y = Contribution)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_flip() +
  labs(title = "Contribution of Variables to Dim 1", x = "", y = "Contribution (%)") +
  theme_minimal()

p2 = ggplot(contrib_PC2, aes(x = reorder(Variable, Contribution), y = Contribution)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_flip() +
  labs(title = "Contribution of Variables to Dim 2", x = "", y = "Contribution (%)") +
  theme_minimal()

grid.arrange(p1, p2, ncol=2)
```



```
PC1 <- data.pca$ind$coord[, 1]
PC2 <- data.pca$ind$coord[, 2]
```

```
pca_scores <- as.data.frame(data.pca$ind$coord)
pca_scores$Team <- data$Team
plot_list <- list()
p1 <- ggplot(pca_scores, aes(x = Team, y = Dim.1)) +
        geom_bar(stat = "identity", fill = "skyblue") +
        labs(title = "Scores on PC1 for Each Team", x = "Team", y = "Score on PC1") +
        theme(axis.text.x = element_text(angle = 90, hjust = 1))
p2 <- ggplot(pca_scores, aes(x = Team, y = Dim.2)) +
        geom_bar(stat = "identity", fill = "salmon") +
        labs(title = "Scores on PC2 for Each Team", x = "Team", y = "Score on PC2") +
        theme(axis.text.x = element_text(angle = 90, hjust = 1))
grid.arrange(p1, p2, ncol = 2)
```



- Analyzing the scores on the first principal component (PC1) of the La Liga teams helps us understand the main distinctions between the teams based on the input variables used in the PCA.
    - **Teams with high scores on PC1**: Athletic, Atletico Madrid, Barcelona, Real Marid: These teams may have very different characteristics and performance that stand out from the other teams in the original data. They may have better records,more assistances ,more wins,more successful passes and make more dynamic attacking transitions.

  - **Teams with low scores on PC1**: Espanyol, Getafe, Granada, Levante, Sporting Gijon: These teams have significantly negative scores on PC1. This suggests that they may have the opposite characteristics to high scoring teams. They may have poorer performance, poorer records, or lower performance metrics and tend to commit more fouls.
  - **Teams with negative scores but not too low on PC1**: They may have below average performance, they may struggle in competition, but they are not the worst teams.
- While PC1 explains the overall performance of teams, PC2 focuses on variables that reflect the team's defensive performance, including the number of matches without conceding a goal, the number of goals conceded, the ratio of goals conceded to total goals conceded, the number of red cards, and the number of offsides.
  - **Teams with high scores on PC2**: Rayo Vallecano, Real Madrid, Granada,..: Teams with high PC2 scores typically exhibit strong defensive records, characterized by numerous clean sheets and a tendency to concede relatively few goals. Their disciplined approach is evident in the lower incidence of red cards, reflecting a commitment to maintaining defensive stability. Moreover, these teams often adopt a tactically aggressive style of play, resulting in a higher number of offsides for the opposition. This proactive defensive approach helps in disrupting opponent attacks and controlling the flow of the game.

  - **Teams with low scores on PC2**: Teams with low PC2 scores typically demonstrate weaker defensive records, characterized by fewer clean sheets and a tendency to concede more goals. Their poor discipline is evident in higher red card counts, indicating a lack of control and organization on the defensive end. Additionally, these teams often employ less aggressive tactics, leading to fewer offsides for the opposition. This defensive approach may lack proactive measures to disrupt opponent attacks effectively, resulting in a higher vulnerability to conceding goals.

  - **Specific to Atlético Madrid:** Reason for Low PC2 Score: Despite many clean sheets, other factors such as a higher number of fouls and red cards negatively impact their PC2 score. Their aggressive defensive style might result in fewer offsides, contributing to a lower score.

3. Discussion ()

- In this analysis, we have explored the Olympic dataset and performed a Principal Components Analysis (PCA) to identify the underlying structure of the data. We found that the first three principal components explain 80.3% of the total variance, which is a good starting point for further analysis. We visualized the data in a 3D plot and identified the main variables that contribute to each principal component. We also visualized the scores of each nation on the first two principal components. This analysis provides valuable insights into the performance of different nations in the Olympic games and can help identify patterns and trends in the data.
- The results showed that the best teams were characterized and differentiated from the bottom teams by completing more successful passes and making more dynamic attacking transitions. The bottom teams were characterized by making more defensive than attacking moves, scoring fewer goals and spending more time in the final third of the pitch.