DATA 607 Assignment 1

Warner Alexis

2023-09-06

Introduction

I select the SPI Rating for Football club Soccer data set. SPI defined as Soccer Power Index which is a rating system designed to rank Soccer Clubs' overall strength. In addition, this is rating system that also use to designate the best team status based on their offensive and attacking strength. This data set contains 641 club teams ranking from 1(the best) to 641(the worst) with their average offensive and defensive rate per game and the team overall psi rating. The link to retrieve the data is below: <https://projects.fivethirtyeight.com/soccer-api/club/spi_global_rankings.csv (https://projects.fivethirtyeight.com/soccer-api/club/spi_global_rankings.csv).

Including Plots

You can also embed plots, for example:

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────── tidyverse 2.0.0 ──
## ✓ dplyr      1.1.2      ✓ readr      2.1.4
## ✓ forcats    1.0.0      ✓ stringr    1.5.0
## ✓ ggplot2    3.4.2      ✓ tibble     3.2.1
## ✓ lubridate  1.9.2      ✓ tidyr      1.3.0
## ✓ purrr      1.0.1
## ── Conflicts ──────────────────────────────────── tidyverse_conflicts() ──
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(hrbrthemes)
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.
##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and
##       if Arial Narrow is not on your system, please see https://bit.ly/arialnarrow
```

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(gt)
club <- read.csv("https://projects.fivethirtyeight.com/soccer-api/club/spi_global_rankings.csv",
sep = ',',
                 stringsAsFactors = F)
head(club,10)
```

```
##    rank prev_rank                    name                 league  off  def
## 1     1         1         Manchester City  Barclays Premier League 2.79 0.28
## 2     2         2           Bayern Munich        German Bundesliga 3.04 0.68
## 3     3         3               Barcelona Spanish Primera Division 2.45 0.43
## 4     4         4              Real Madrid Spanish Primera Division 2.56 0.60
## 5     5         5               Liverpool  Barclays Premier League 2.63 0.67
## 6     6         6                  Arsenal  Barclays Premier League 2.53 0.61
## 7     7         7                Newcastle  Barclays Premier League 2.38 0.53
## 8     8         8                   Napoli           Italy Serie A 2.30 0.51
## 9     9         9        Borussia Dortmund        German Bundesliga 2.83 0.84
## 10   10        10 Brighton and Hove Albion  Barclays Premier League 2.47 0.73
##      spi
## 1  92.00
## 2  87.66
## 3  86.40
## 4  84.41
## 5  83.93
## 6  83.92
## 7  83.70
## 8  83.25
## 9  82.91
## 10 80.88
```

```
# Understanding the variables in the data set
summary(club)
```

```
##       rank         prev_rank         name               league
## Min.   : 1   Min.    : 1   Length:641        Length:641
## 1st Qu.:161   1st Qu.:161   Class :character   Class :character
## Median :321   Median :321   Mode  :character   Mode  :character
## Mean   :321   Mean    :321
## 3rd Qu.:481   3rd Qu.:481
## Max.   :641   Max.    :641
##       off             def             spi
## Min.   :0.200   Min.    :0.280   Min.    : 4.86
## 1st Qu.:0.850   1st Qu.:1.180   1st Qu.:26.68
## Median :1.180   Median :1.460   Median :38.88
## Mean   :1.213   Mean    :1.479   Mean    :40.27
## 3rd Qu.:1.530   3rd Qu.:1.760   3rd Qu.:52.11
## Max.   :3.040   Max.    :2.860   Max.    :92.00
```

Lets remove the columns we don't need and rename some of them so it would be easier to understand the data set.

We save the data in our Github and reload as instructed.

link: https://raw.githubusercontent.com/joewarner89/CUNY-607/main/homeworks/Assignment%201/data/spi_global_rankings.csv (https://raw.githubusercontent.com/joewarner89/CUNY-607/main/homeworks/Assignment%201/data/spi_global_rankings.csv)

```
club <- read.csv("https://raw.githubusercontent.com/joewarner89/CUNY-607/main/homeworks/Assignme
nt%201/data/spi_global_rankings.csv",
          stringsAsFactors = F, header = T, sep = ',')

club <- club %>% select(-contains("prev_rank"))
head(club)
```

```
##   rank            name                  league off  def   spi
## 1    1 Manchester City  Barclays Premier League 2.79 0.28 92.00
## 2    2   Bayern Munich        German Bundesliga 3.04 0.68 87.66
## 3    3       Barcelona Spanish Primera Division 2.45 0.43 86.40
## 4    4     Real Madrid Spanish Primera Division 2.56 0.60 84.41
## 5    5       Liverpool  Barclays Premier League 2.63 0.67 83.93
## 6    6         Arsenal  Barclays Premier League 2.53 0.61 83.92
```

```
club <- club %>% rename(club_team = name,
                        offensive_rate = off,
                        defensive_rate = def
                        )
club$power_class <- as.factor(ifelse(club$spi>= .01 & club$spi <= 29.99, 'Worst Rating Team',
                            ifelse(club$spi >= 30 & club$spi <= 39.99, 'Average Team',
                                ifelse(club$spi >= 40 & club$spi <= 75.99 , 'Good Tea
m',
                                        ifelse(club$spi >=76 & club$spi <= 82.99, 'Poten
tial World Class Team',
                                            ifelse(club$spi >=83 & club$spi <= 100,
'World Class Team', 'Unknown'))))))
head(club,10)
```

```
##    rank                 club_team                      league offensive_rate
## 1     1           Manchester City  Barclays Premier League           2.79
## 2     2             Bayern Munich         German Bundesliga           3.04
## 3     3                 Barcelona Spanish Primera Division           2.45
## 4     4               Real Madrid Spanish Primera Division           2.56
## 5     5                 Liverpool  Barclays Premier League           2.63
## 6     6                   Arsenal  Barclays Premier League           2.53
## 7     7                 Newcastle  Barclays Premier League           2.38
## 8     8                    Napoli            Italy Serie A           2.30
## 9     9         Borussia Dortmund         German Bundesliga           2.83
## 10   10 Brighton and Hove Albion  Barclays Premier League           2.47
##    defensive_rate   spi               power_class
## 1            0.28 92.00          World Class Team
## 2            0.68 87.66          World Class Team
## 3            0.43 86.40          World Class Team
## 4            0.60 84.41          World Class Team
## 5            0.67 83.93          World Class Team
## 6            0.61 83.92          World Class Team
## 7            0.53 83.70          World Class Team
## 8            0.51 83.25          World Class Team
## 9            0.84 82.91 Potential World Class Team
## 10           0.73 80.88 Potential World Class Team
```

the variable spi determine the power class of the team. The higher the spi rate the better is the team. Manchester City rank # 1 because it has the highest spi 92.00

The top ten team of 2022-2023 season :

```
# select only top 20 teams based on ordered spi in the data set
top_20 <- head(club,20)
top_20 %>%
  ggplot( aes(x=club_team, y=spi) ) +
  geom_bar(stat="identity", fill="#69b3a2") +
  coord_flip() +
  theme_ipsum() +
  theme(
    panel.grid.minor.y = element_blank(),
    panel.grid.major.y = element_blank(),
    legend.position="none"
  ) +
  xlab("Top 20 Teams Playing World Class Football") +  ggtitle("Highest Rated Soccer Clubs")+
  ylab("Soccer Power Index(Best Team for 2022-2023 season)")
```

Highest Rated Soccer Clubs
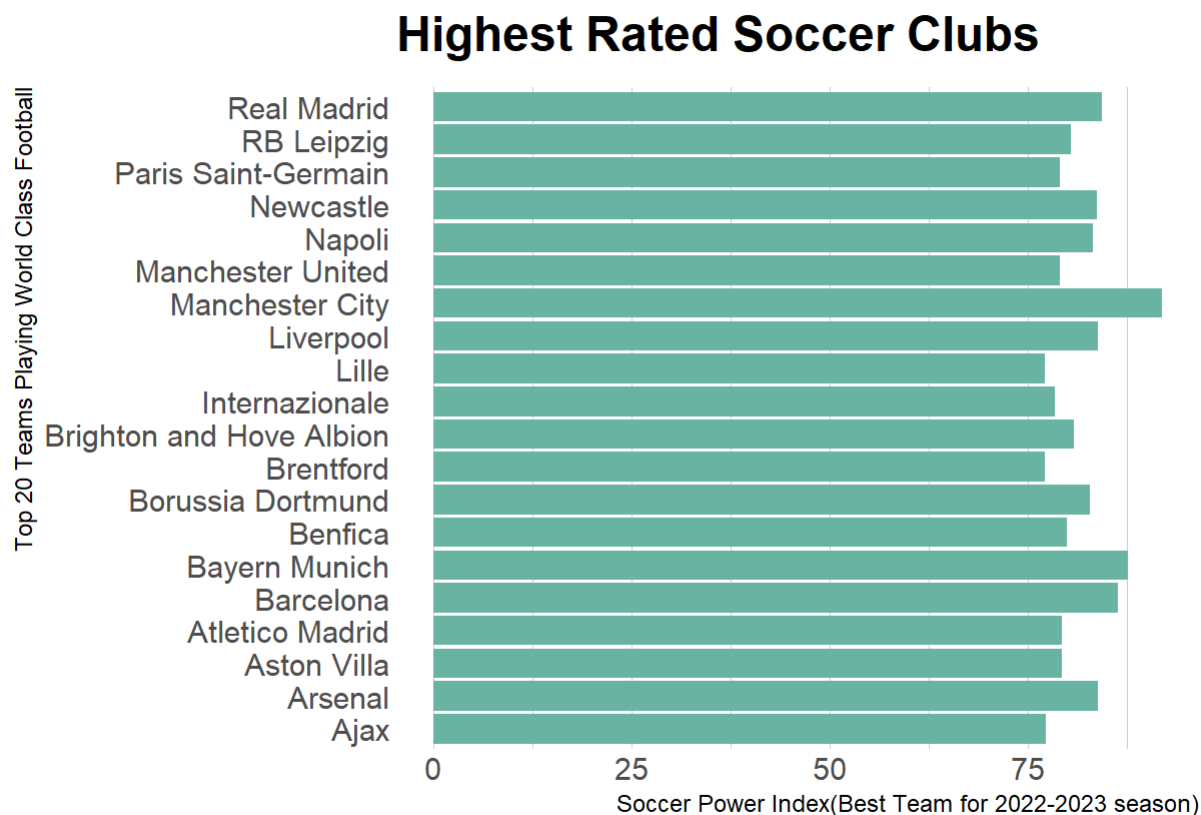

Lets see the 20 teams with the lowest rating

```
worst_20 <- tail(club,20)
worst_20 %>%
  ggplot( aes(x=club_team, y=spi) ) +
  geom_bar(stat="identity", fill="#69b3a2") +
  coord_flip() +
  theme_ipsum() +
  theme(
    panel.grid.minor.y = element_blank(),
    panel.grid.major.y = element_blank(),
    legend.position="none"
  ) +
  xlab("Top 20 Teams Playing Poorly") +  ggtitle("Lowest Rated Soccer Club")+
  ylab("Soccer Power Index(Best Team for 2022-2023 season)")
```
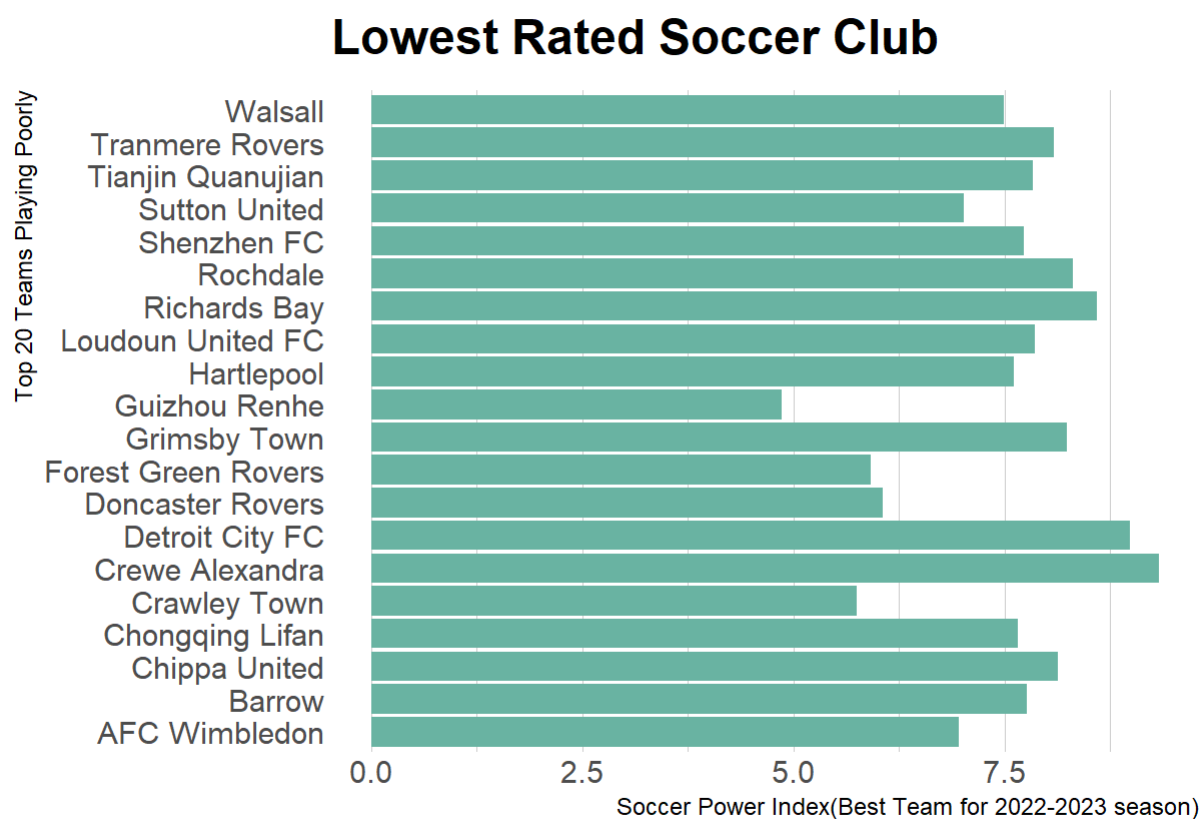

During the 2022-2023 season, a lot team improved their goal per game ratio and reduced conceding goals.SPI data set has the overall estimate for All top Soccer Clubs in the world. Lets explore the relationships between the variables.

```
# best offensive teams and defensive teams
off <-
  club %>% arrange(desc(offensive_rate)) %>% head(10) %>% select(club_team, offensive_rate)

# Top 10 Offensive Team
gt(off) %>%
  tab_header(
    title = "Best Offensive Team for the 2022-2023 Season",
    subtitle = "Highest scoring Team for 2022-2023 Season "
  )
```

Best Offensive Team for the 2022-2023 Season

Highest scoring Team for 2022-2023 Season

club_team	offensive_rate
Bayern Munich	3.04
Borussia Dortmund	2.83
Manchester City	2.79
Ajax	2.66
Liverpool	2.63
Paris Saint-Germain	2.62
Real Madrid	2.56
Arsenal	2.53
Celtic	2.53
Brighton and Hove Albion	2.47

```
# Best Defense in Europe
deff <-
  club %>% arrange(defensive_rate) %>% head(10) %>% select(club_team, defensive_rate)

# Top 10 Defensive Team
gt(deff) %>%
  tab_header(
    title = "Best Defensive Team for the 2022-2023 Season",
    subtitle = "Highest scoring Team for 2022-2023 Season "
  )
```

Best Defensive Team for the 2022-2023 Season

Highest scoring Team for 2022-2023 Season

club_team	defensive_rate
Manchester City	0.28
Barcelona	0.43
Napoli	0.51
Aston Villa	0.51
Real Sociedad	0.52
Newcastle	0.53
Athletic Bilbao	0.59
Real Madrid	0.60
Arsenal	0.61
Crystal Palace	0.62

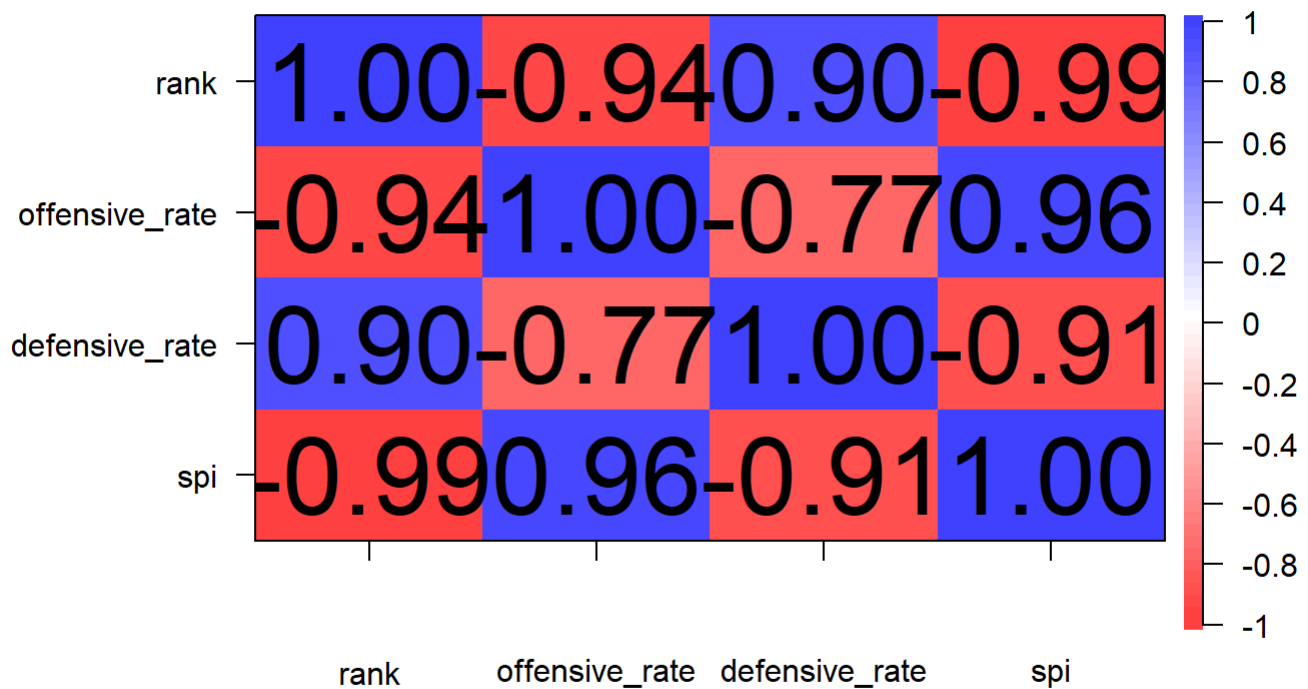Lets look at the relationship between these variables.

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
# creating only numerical variables
corr <- club %>% select(rank,offensive_rate,defensive_rate,spi)
corPlot(corr[,1:4], main = "Correlation of Team Statistic")
```

Correlation of Team Statistic


```
corres <- cor(corr)
corres <- round(corres, 2)
# Transform the correlation table to data frame before using gt pckg
gt(data.frame(round(corres,2))) %>%
  tab_header(
    title = "Correlation Of All the Features for the 2022-2023 Season",
    subtitle = " Relationship of All Soccer Statistics"
  )
```

Correlation Of All the Features for the 2022-2023 Season

Relationship of All Soccer Statistics

rank	offensive_rate	defensive_rate	spi
1.00	-0.94	0.90	-0.99
-0.94	1.00	-0.77	0.96
0.90	-0.77	1.00	-0.91
-0.99	0.96	-0.91	1.00

Conclusion:

What we learn from the data is that a team cannot win a tournament without a good defense and defensive rate is highly correlated with the ranking number 1. Manchester City conceived few goals than any other teams in Europe. They have won UEFA Champion League, Premier League, FA Cup, Community Shield and EUFA Super Cup. Soccer Power Index represents the team's overall strength over 100. SPI is a mixture of both defensive and offensive ratings. The team with the highest SPI would occupy the rank 1 as best team in the world.