

Second Season More Difficult

Umid Kumar Dey

3/11/2021

This document contains the code for the model that tests whether the saying that the second season is the hardest for newly-promoted teams is true or not.

```
# Plot the Data
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
library(tidyr)
library(reshape)
```

```
##
## Attaching package: 'reshape'

## The following objects are masked from 'package:tidyr':
##
##   expand, smiths

## The following object is masked from 'package:dplyr':
##
##   rename
```

```
# The dataset was downloaded from ESPN using a scraper and had to be
# cleaned/processed before it could be used.
```

```
df_cleaner <- function(df) {
  d1<-data.frame(df[])
  d2 <- subset(d1, select = -c(1,3, 5, 6, 8, 9, 11, 12, 14, 15))
  colnames(d2) <- c('Team', 'position.1st.season', 'position.2nd.season', 'position.3rd.season', 'positi
```

```

d2$Team <- make.unique(d2$Team)

return(d2)
}

# The graph plotter function plots the progress graph of newly promoted teams.
# Since Bundesliga has only 18 teams in the league, compared to 20 that the
# other leagues in the top 5 have, the graph plotter for it is slightly
# different.

buli_graph_plotter <- function(df) {
  d3 <- df_cleaner(df)
  lab <- c(sum(is.na(d3$position.2nd.season)),sum(is.na(d3$position.3rd.season)), sum(is.na(d3$position
  colnames(d3)<-c('Team', 'x1', 'x2', 'x3', 'x4')
  df_clean <- d3 %>%
    gather(season, value, num_range('x', 1:4)) %>%
    mutate(season = readr::parse_number(season))

  plots <- ggplot(df_clean, aes(x = season, y = value, colour = Team)) +
    geom_line() + ylab('Position in the league') +
    scale_y_continuous(trans = 'reverse', breaks = 0:20) +
    stat_summary(aes( y = value), fun=mean, colour="black", geom="line",group=1, size = 1) +
    scale_color_manual(values = c(gray.colors(nrow(df)))) + theme(legend.position="none") +
    geom_hline(yintercept = 16, color = 'Red', linetype = 'dashed', size = 1)+
    scale_x_continuous(
      name = "Number of relegated teams",
      labels = lab,
      sec.axis = sec_axis( trans=~.*1, name="Season number")
    )

  return(plots)
}

graph_plotter <- function(df) {

  d3 <- df_cleaner(df)
  lab <- c(sum(is.na(d3$position.2nd.season)),sum(is.na(d3$position.3rd.season)), sum(is.na(d3$position

  colnames(d3)<-c('Team', 'x1', 'x2', 'x3', 'x4')
  means <- cbind(mean(na.omit(d3$x2)), mean(na.omit(d3$x1)), mean(na.omit(d3$x3)), mean(na.omit(d3$x4)))
  df_clean <- d3 %>%
    gather(season, value, num_range('x', 1:4)) %>%
    mutate(season = readr::parse_number(season))

  plots <- ggplot(df_clean, aes(x = season, y = value, color = Team)) +
    geom_line() + ylab('Position in the league') + scale_y_continuous(trans = 'reverse', breaks = 0:20)
    stat_summary(aes( y = value), fun=mean, colour="black", geom="line",group=1, size = 1)+
    scale_color_manual(values = c(gray.colors(nrow(df)))) + theme(legend.position="none") +
    geom_hline(yintercept = 18, color = 'Red', linetype = 'dashed', size = 1) +
    scale_x_continuous(

```

```

    name = "Number of relegated teams",
    labels = lab,
    sec.axis = sec_axis( trans=~.*1, name="Season number")
  )

  return(plots)
}

# Get Data
library(readxl)
buli <- read_excel("E:/TUD/Course materials/Sports Data Visualisation/BuLi/Relegation topic/Daten Aufst

## New names:
## * ' -> ...1
## * 'relegated?' -> 'relegated?...6'
## * 'relegated?' -> 'relegated?...9'
## * 'relegated?' -> 'relegated?...12'
## * 'relegated?' -> 'relegated?...15'

liga<- read_excel("E:/TUD/Course materials/Sports Data Visualisation/LaLiga/Relegation topic/Daten Aufst

## New names:
## * ' -> ...1
## * 'relegated?' -> 'relegated?...6'
## * 'relegated?' -> 'relegated?...9'
## * 'relegated?' -> 'relegated?...12'
## * 'relegated?' -> 'relegated?...15'

seria <- read_excel("E:/TUD/Course materials/Sports Data Visualisation/Serie A/Relegation topic/Serie A

## New names:
## * ' -> ...1
## * 'relegated?' -> 'relegated?...6'
## * ' -> ...7
## * 'relegated?' -> 'relegated?...10'
## * ' -> ...11
## * ...

seria <- subset(seria, select = -c(7, 11, 15))

# The dataset for Ligue 1 was slightly differently stored, which is why it had
# to be edited slightly to fit the data cleaner function

ligue <- read_excel("E:/TUD/Course materials/Sports Data Visualisation/Ligue1/Relegation topic/Ligue 1

## New names:
## * 'relegated?' -> 'relegated?...5'
## * 'relegated?' -> 'relegated?...8'

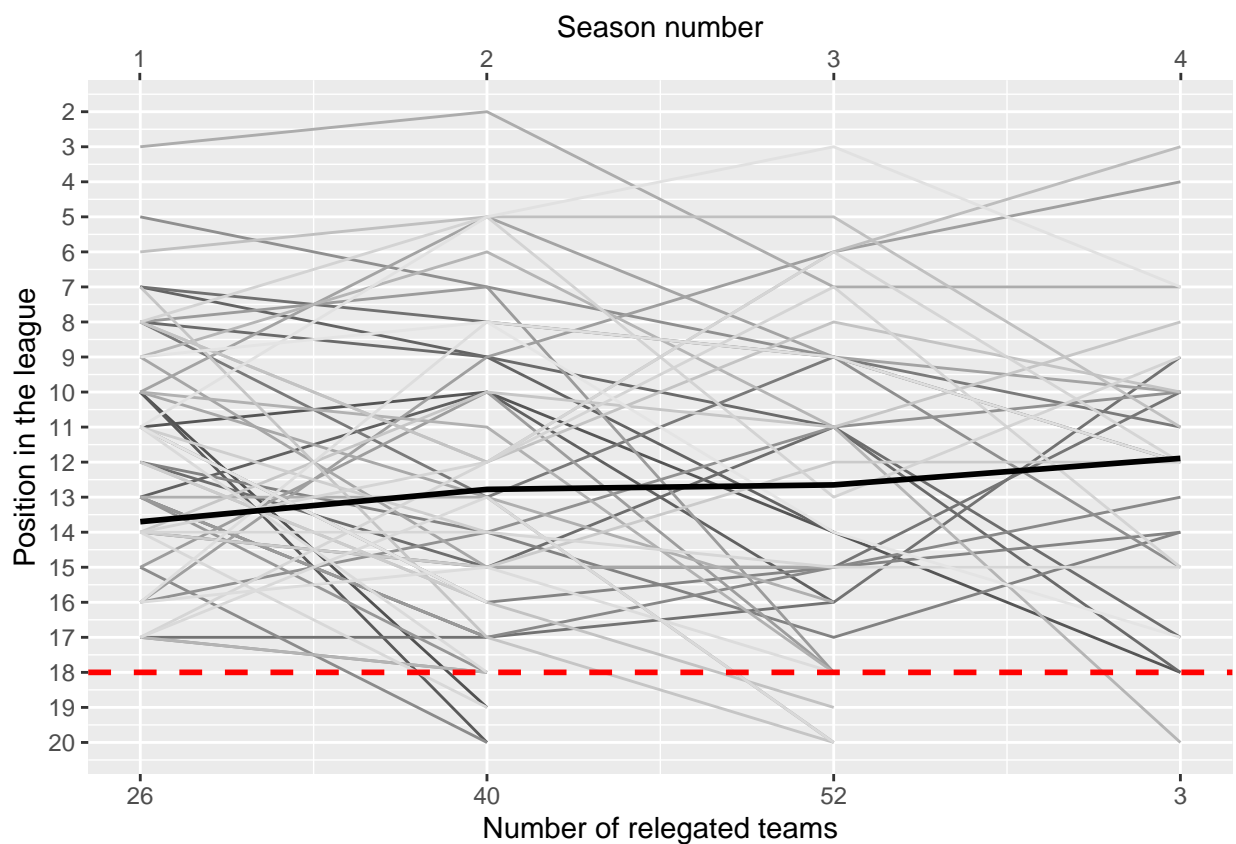
```

```
## * 'relegated?' -> 'relegated?...11'
## * 'relegated?' -> 'relegated?...14'
## * 'relegated?' -> 'relegated?...17'
## * ...
```

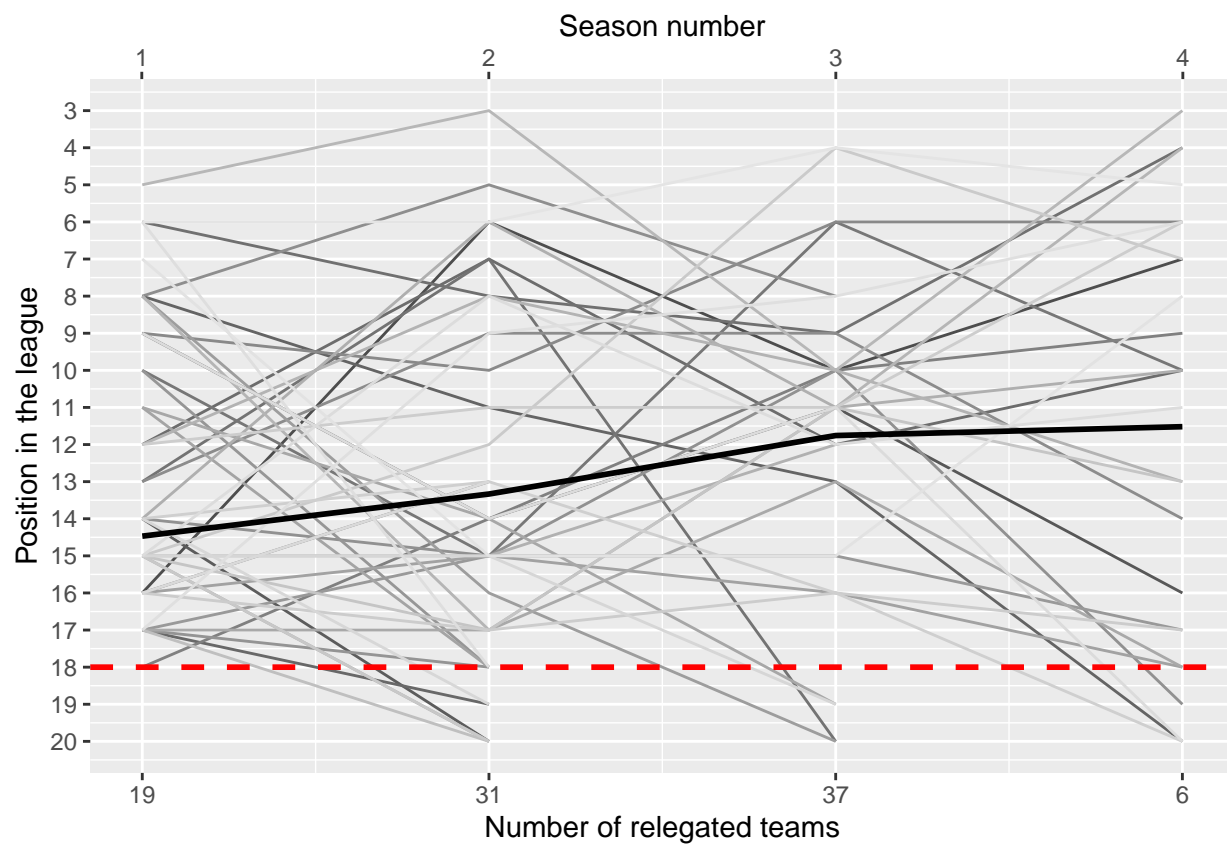
```
ligue <- cbind(1, ligue)
ligue <- ligue[-c(1:19), ]
ligue <- subset(ligue, select = -seq(16,20))
epl <- read_excel("E:/TUD/Course materials/Sports Data Visualisation/EPL/Relegation topic/Daten Aufsteig
```

```
## New names:
## * ' ' -> ...1
## * 'relegated?' -> 'relegated?...6'
## * 'relegated?' -> 'relegated?...9'
## * 'relegated?' -> 'relegated?...12'
## * 'relegated?' -> 'relegated?...15'
```

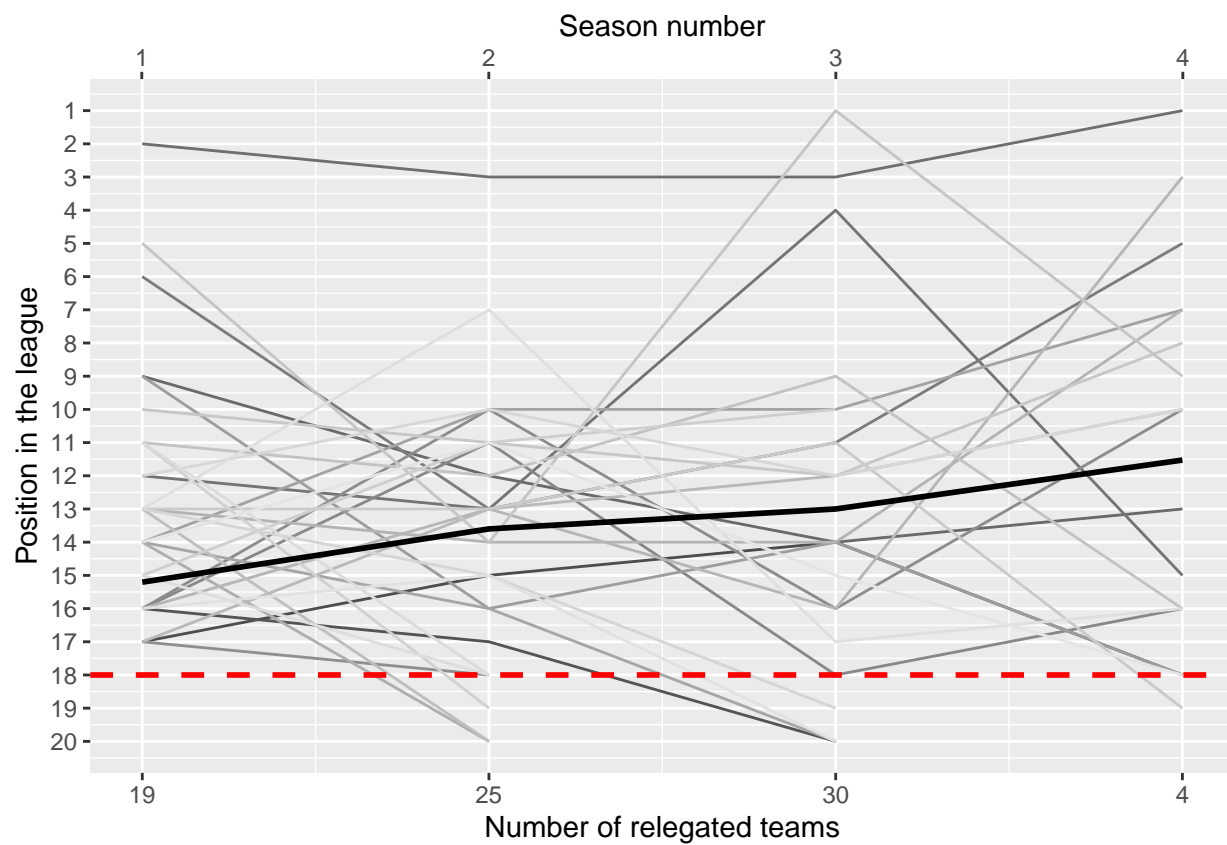
```
graph_plotter(seria)
```



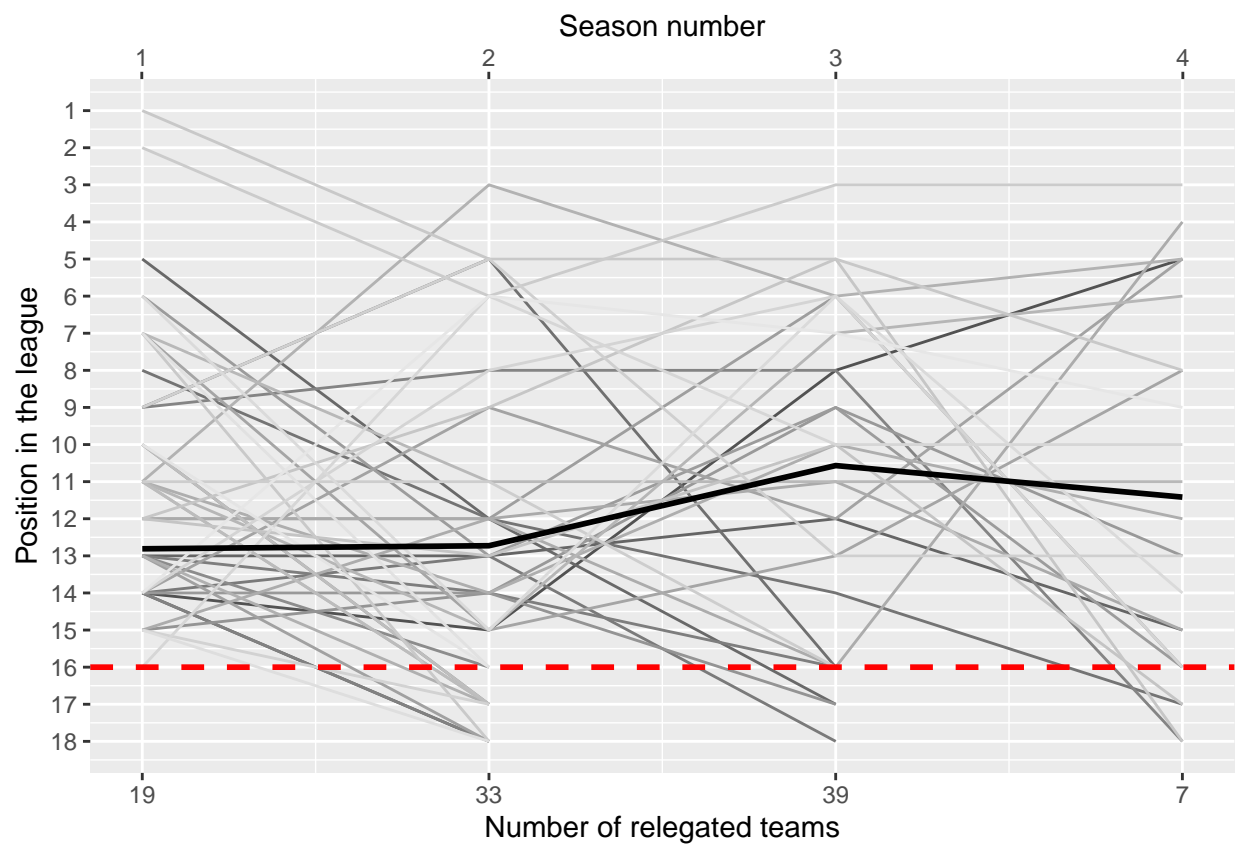
```
graph_plotter(liga)
```



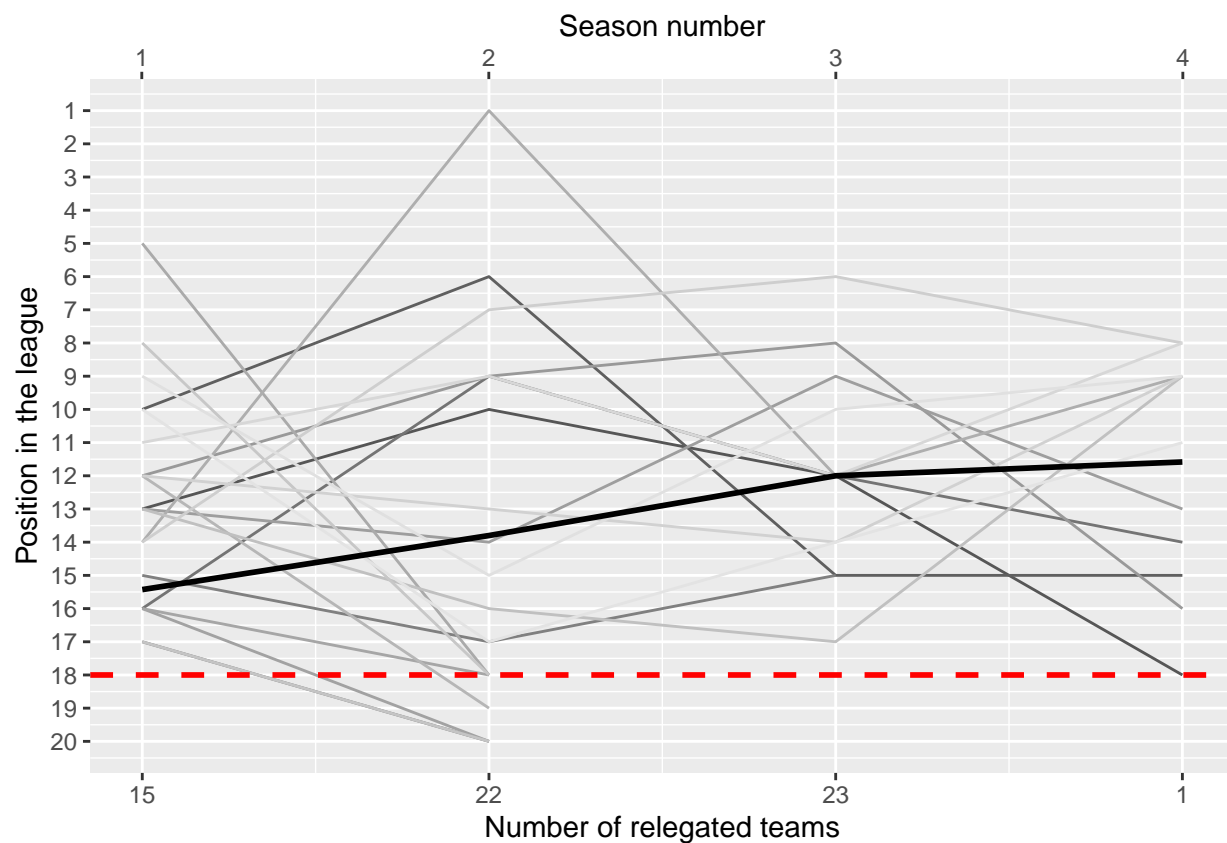
```
graph_plotter(ligue)
```



```
buli_graph_plotter(buli)
```



```
graph_plotter(ep1)
```



The darker line is the mean line for all the newly promoted teams. As can be seen from the graphs, second season being harder might truly be a myth because in most cases, the first season was the hardest for newly-promoted teams