FIFA soccer data 2019

Jeel Bhalodia, Saurav Kumar, Silke Meiner

Soccer may be a game to some. It reflects society and can be taken seriously. We present three cases.

1. / In chapter 1 we investigate the distribution of wealth within European clubs. We find that wealth is not evenly shared between clubs. A club with players of higher values is often seen in a leading final position in its league.
2. / In chapter 2 we follow the migration of players from poor countries to rich. Brasil takes a special role as a country most unable to feed its talented players into its own clubs.
3. / In chapter 3 we evaluate players performance on field positions and compare teams wrt their players performance. We present convincing visualisations indicating that a clubs final position in its league is related to its players performance in the field positions. This seems to hold true across European premier leagues.

```
library(ggplot2)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
##
## 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
```

```
tinytex::install_tinytex()
```

```
## TinyTeX installed to /Users/silke/Library/TinyTeX
```

Distribution of wealth in Bundesliga

At the German premier league called Bundesliga, wealth is distributed unevenly between clubs.

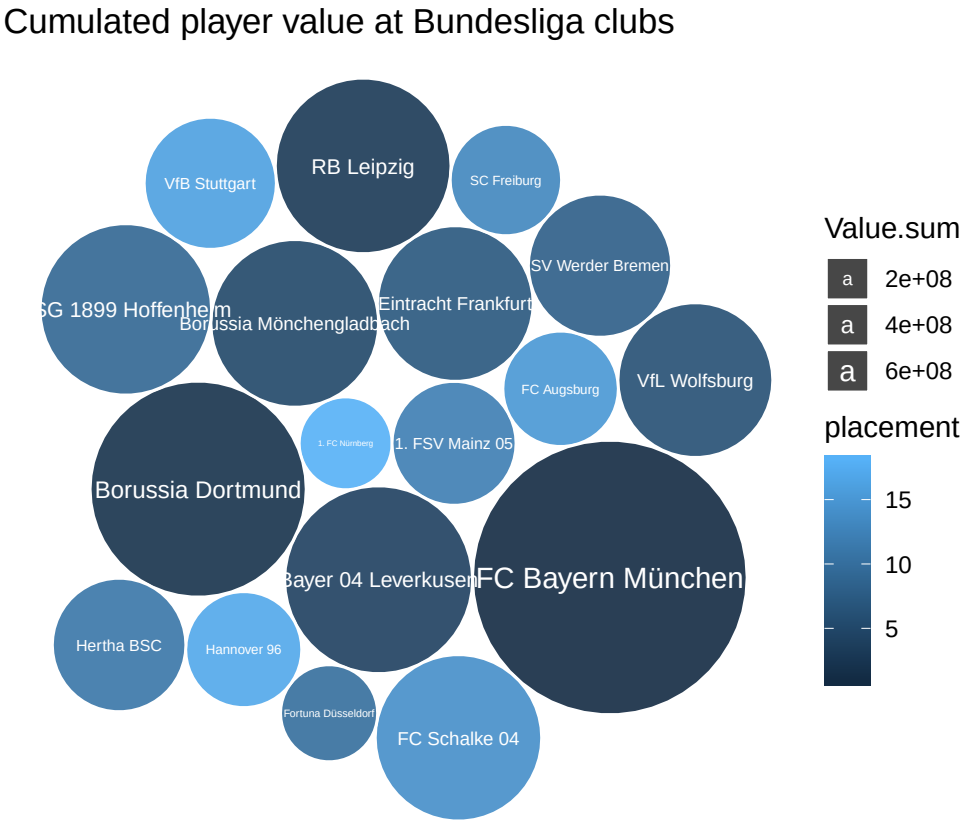
We show a circles packaging where the size of a circle associated with a club corresponds to the cumulated value of its players. Large circles represent rich clubs. Additional information is put in the color of the circles: Dark colours corresponds to a final placement at the top of the league, whereas light colours correspond to a placement at the bottom of the league.

```r
library(packcircles)

# read data
soccer.1 <- read.csv("soccer-preprocessed-01.csv")
final.placement <- read.csv("abschlusstabelle.csv") %>%
  select(Tabellenplatz, club)
```

```r
## clubs from bundesliga
BL_Clubs <- final.placement$club
value.df <- filter(soccer.1, Club %in% BL_Clubs) %>%
  group_by(Club) %>%
  summarise(Value.sum=sum(Value)) %>%
  data.frame()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
value.df <- merge(value.df , final.placement, by.x= 'Club', by.y='club')

packing <- circleProgressiveLayout(value.df$Value.sum, sizetype ='area')

# add packing information to the data frame
value.df <- cbind(value.df, packing)

npoints<-50
dat.gg <- circleLayoutVertices(packing, npoints=npoints)
dat.gg <- cbind(dat.gg, rep(value.df$Tabellenplatz,each=npoints+1))
names(dat.gg)[4] <- 'placement'

# Make the plot
p <- ggplot() +
  ggtitle("Cumulated player value at Bundesliga clubs") +

  # Make the bubbles
  geom_polygon(data = dat.gg, show.legend = TRUE , aes(x, y, group = id, fill = placement), colour = "wh

  # Add text in the center of each bubble + control its size
  geom_text(data = value.df, colour = "white",  aes(x, y, size=Value.sum, label = Club)) +
  scale_size_continuous(range = c(1,4)) +

  # General theme:
  theme_void() +
  theme(legend.position="right") +
  coord_equal()

show(p)
```

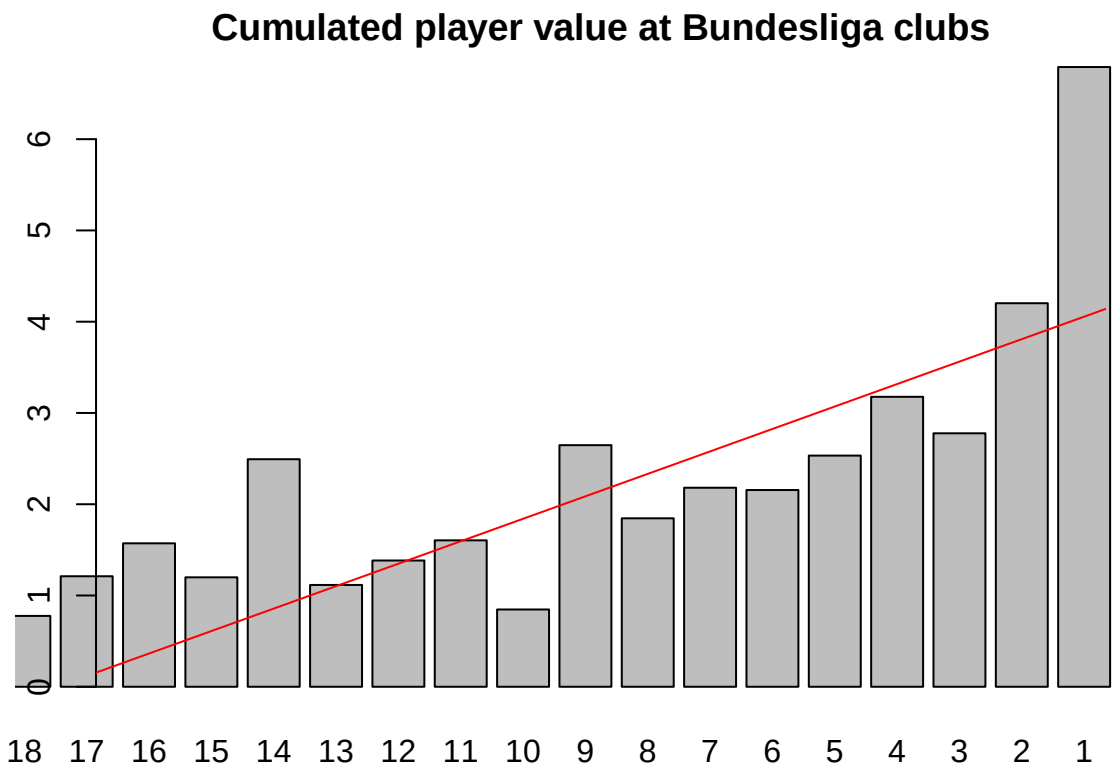


The visualisation implies that clubs with a higher cumulated value of its players placed higher and were more successful.

Room for improvemet: 1. Larger plot to get club labels inside the circles. Otherwise as white on white not readable. 2. the placement sccale should be reversed. Placing as firs should be on the top of the scale / legend. 3. in legend: Rename Value.sum as value. The title explains what exactly the value is. 4. in legend: use decimal, not scientific number coding. Or rather 2 bn € instead of 2e+08. Also adding the € as information

To investigate this further we would visualize the value of a club as bar charts, ordering clubs as in heir final placement. Also a regression could be more exact.

```
par(mfrow=c(1,1), mar=c(2,2,2,2)+0.1 , oma=c(1,1,1,1))
bp <- barplot(value.df$Value.sum / 1e+08 ~ value.df$Tabellenplatz, xlim = rev(range(1:19)) , main="Cumul
lm.obj<- lm(value.df$Value.sum / 1e+08 ~ value.df$Tabellenplatz)
#lm.obj$coeff
abline(lm.obj$coeff, col='red')
```



Room for improvement: 1. The bars should begin at the origin with x=0 not x=-1.5 2. colouring should be the same as in the plot above, so make it a ggplot. 3. a legend is needed to see which bar corresponds to which club. So one can compare the vague impression from the circles plot to the exact information in the barchart.