The Beauty of Soccer

yining wang

2021/4/29

About the data set

The football data set provides a granular view of 9,074 games, totaling 941,009 events from the biggest 5 European football (soccer) leagues: England, Spain, Germany, Italy, France from 2011/2012 season to 2016/2017 season as of 25.01.2017. The first one is a 941,009 * 22 data frame which each row represents an event like attempt, goal, foul, penalty in one football match. The columns are about the features of this event, including players, times, location, description and so on. The second one is a 10,112 * 18 matrix which each row represents one match. The columns are the features of this match, like date, teams, scores, odds in gambling, etc.

Most of the features of events are expressed by numbers. The details are contained in 'dict'.

```
glimpse(events)
```

```
## Rows: 941,009
## Columns: 22
## $ id_odsp       <chr> "UFot0hit/", "UFot0hit/", "UFot0hit/", "UFot0hit/", "...
## $ id_event      <chr> "UFot0hit1", "UFot0hit2", "UFot0hit3", "UFot0hit4", "...
## $ sort_order    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ time          <dbl> 2, 4, 4, 7, 7, 9, 10, 11, 11, 13, 14, 14, 14, 17, 19,...
## $ text          <chr> "Attempt missed. Mladen Petric (Hamburg) left footed ...
## $ event_type    <dbl> 1, 2, 2, 3, 8, 10, 2, 8, 3, 3, 8, 1, 3, 1, 1, 3, 8, 1...
## $ event_type2   <dbl> 12, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 12, NA, 1...
## $ side          <dbl> 2, 1, 1, 1, 2, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 1, 2, 2,...
## $ event_team    <chr> "Hamburg SV", "Borussia Dortmund", "Borussia Dortmund...
## $ opponent      <chr> "Borussia Dortmund", "Hamburg SV", "Hamburg SV", "Ham...
## $ player        <chr> "mladen petric", "dennis diekmeier", "heiko westerman...
## $ player2       <chr> "gokhan tore", "dennis diekmeier", "heiko westermann"...
## $ player_in     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ player_out    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ shot_place    <dbl> 6, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 13, NA, 4,...
## $ shot_outcome  <dbl> 2, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 2, NA, 1, ...
## $ is_goal       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,...
## $ location      <dbl> 9, NA, NA, NA, 2, NA, NA, 2, NA, NA, 4, 15, NA, 9, 15...
## $ bodypart      <dbl> 2, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 1, NA, 2, ...
## $ assist_method <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0,...
## $ situation     <dbl> 1, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 1, NA, 1, ...
## $ fast_break    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

```
glimpse(ginf)
```

```
## Rows: 10,112
## Columns: 18
## $ id_odsp   <chr> "UFot0hit/", "Aw5DflLH/", "bkjpaC6n/", "CzPV312a/", "GUOd...
## $ link_odsp <chr> "/soccer/germany/bundesliga-2011-2012/dortmund-hamburger-...
## $ adv_stats <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRU...
## $ date      <date> 2011-08-05, 2011-08-06, 2011-08-06, 2011-08-06, 2011-08-...
## $ league    <chr> "D1", "D1", "D1", "F1", "F1", "D1", "F1", "F1", "F1", "D1...
```

```
## $ season    <dbl> 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 201...
## $ country   <chr> "germany", "germany", "germany", "france", "france", "ger...
## $ ht        <chr> "Borussia Dortmund", "FC Augsburg", "Werder Bremen", "Par...
## $ at        <chr> "Hamburg SV", "SC Freiburg", "Kaiserslautern", "Lorient",...
## $ fthg      <dbl> 3, 2, 2, 0, 1, 0, 2, 0, 1, 0, 1, 3, 3, 2, 2, 1, 1, 2, 0, ...
## $ ftag      <dbl> 1, 2, 0, 1, 0, 1, 2, 2, 3, 3, 1, 1, 0, 1, 2, 2, 5, 0, 1, ...
## $ odd_h     <dbl> 1.56, 2.36, 1.83, 1.55, 2.50, 2.06, 2.29, 2.80, 4.50, 3.0...
## $ odd_d     <dbl> 4.41, 3.60, 4.20, 4.50, 3.40, 3.75, 3.25, 3.10, 3.55, 3.8...
## $ odd_a     <dbl> 7.42, 3.40, 4.80, 9.40, 3.45, 3.95, 3.85, 3.05, 2.00, 2.5...
## $ odd_over  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ odd_under <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ odd_bts   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ odd_bts_n <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
```

```
head(dict)
```

```
##   event_type        description
## 1          0       Announcement
## 2          1            Attempt
## 3          2             Corner
## 4          3               Foul
## 5          4        Yellow card
## 6          5 Second yellow card
```

data preparation

Since football lottery is not what we are interested in, I just delete that columns. Then, I add two columns presenting the scores of home team and away team in 'ginf' tibble. The criterion is 3 scores for win, 1 score for draw and 0 for lose.

```
# delete odds data
ginf <- ginf[, -c(12:18)]

# join 2 tibbles
events <- left_join(events, ginf)
```

```
## Joining, by = "id_odsp"
```

```
# create 2 columns presenting the scores of home team and away team
ginf$htscore <- NULL
ginf$atscore <- NULL

# 3 scores if win, 1 if draw, 0 if lose
for (i in 1:nrow(ginf)){
  if (ginf$fthg[i] > ginf$ftag[i]){
    ginf$htscore[i] = 3
    ginf$atscore[i] = 0
  }
  if (ginf$fthg[i] == ginf$ftag[i]){
    ginf$htscore[i] = 1
    ginf$atscore[i] = 1
  }
  if (ginf$fthg[i] < ginf$ftag[i]){
```

```
    ginf$htscore[i] = 0
    ginf$atscore[i] = 3
  }
}
```

Who scored the most goals?

The most exciting part of soccer is definitely the goal. Therefore, the first thing I want to detect in this data is that who scored the most goals? ‘table 1’ is a table containing the most efficient shooters who at least shot 200 times according to the goal percentage. The first one is miroslav klose. The total goal number plot reflect the most productive shoots. Unsurprisingly, Messi and Ronaldo are one the top of this plot.

```
# detect the players getting most goals
goal_efficiency <- filter(events, is_goal == 1) %>% group_by(player) %>% summarize(goal.num
ber = n()) %>% arrange(desc(goal.number))

# add 2 columns which are shot numbers and on target shot numbers to tibble 'goal_efficienc
y'
goal_efficiency <- filter(events, event_type == 1) %>% group_by(player) %>% summarize(shot.
number = n()) %>% right_join(goal_efficiency)
```

```
## Joining, by = "player"
```

```
goal_efficiency <- filter(events, shot_outcome == 1) %>% group_by(player) %>% summarize(ont
arget.number = n()) %>% right_join(goal_efficiency)
```

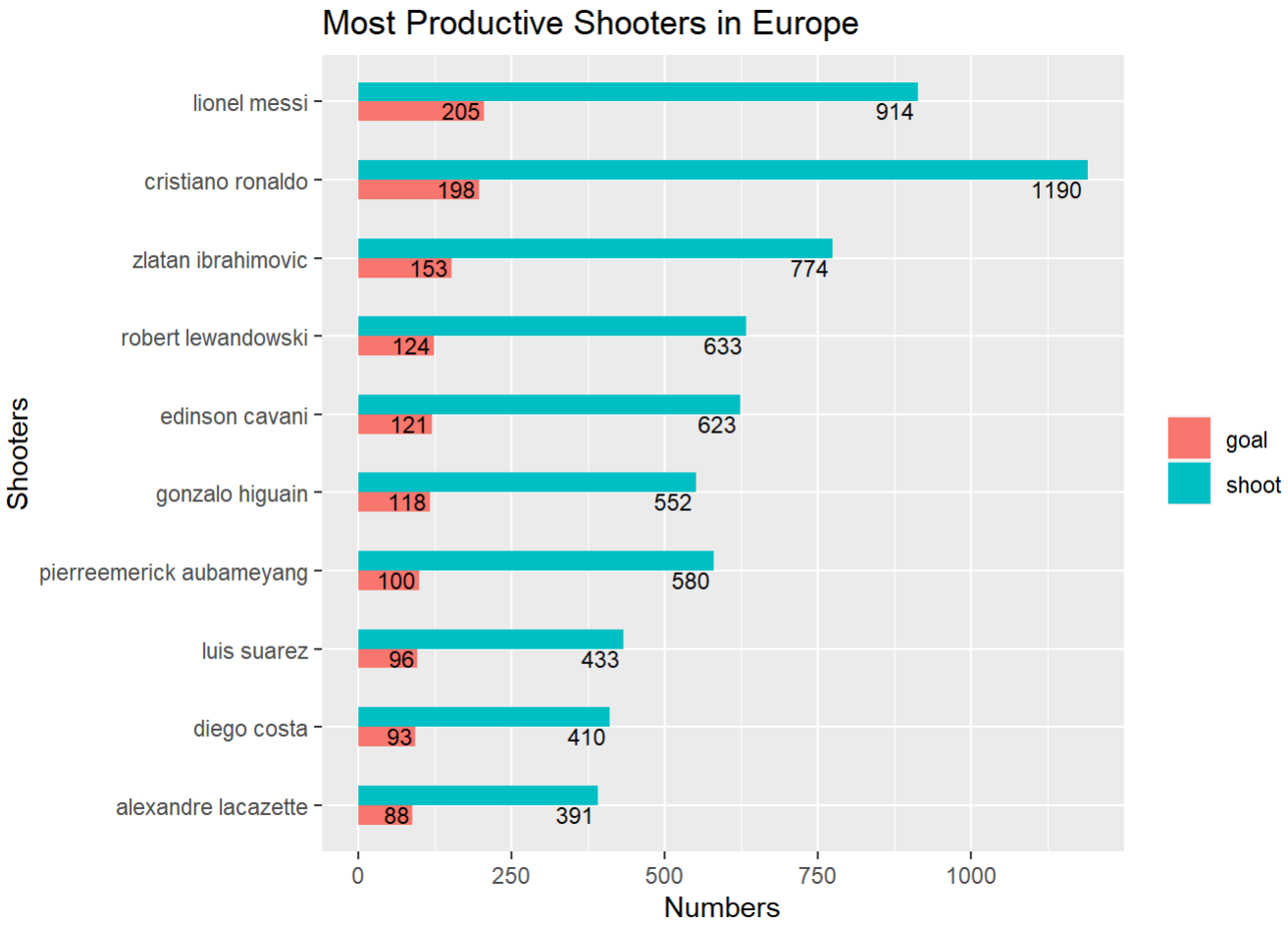
```
## Joining, by = "player"
```

```
# a table containing the most efficient shooters who at least shot 200 times according to t
he goal percentage
goal_efficiency <- goal_efficiency %>% mutate(on.target.percentage = round(ontarget.number/
shot.number, digits = 3), goal.percentage = round(goal.number/shot.number, digits = 3))
table1 <- head(filter(goal_efficiency, shot.number > 200) %>% arrange(desc(goal.percentage)
), n = 10)
table1
```

```
## # A tibble: 10 x 6
##    player ontarget.number shot.number goal.number on.target.perce~
##    <chr>            <int>       <int>       <int>            <dbl>
##  1 miros~              91         205          51            0.444
##  2 carlo~             113         230          57            0.491
##  3 diego~             191         410          93            0.466
##  4 alexa~             193         391          88            0.494
##  5 lione~             435         914         205            0.476
##  6 mauro~             122         323          72            0.378
##  7 luis ~             201         433          96            0.464
##  8 mario~              98         221          48            0.443
##  9 gonza~             261         552         118            0.473
## 10 falcao             171         381          80            0.449
## # ... with 1 more variable: goal.percentage <dbl>
```

```
# plot
```

```
data.for.plot <- pivot_longer(head(arrange(goal_efficiency, desc(goal.number)), n = 10), 3:
4, names_to = 'type', values_to = 'number')
ggplot(data = data.for.plot, aes(x = number, y = player)) + geom_bar(aes(fill = type), stat
 = "identity", position = "dodge", width=.5) + geom_text(aes(label = number), hjust = 1.1,
vjust = 1, size = 3) + scale_y_discrete(limits = rev(c("lionel messi", "cristiano ronaldo",
 "zlatan ibrahimovic", "robert lewandowski", "edinson cavani", "gonzalo higuain", "pierreem
erick aubameyang", "luis suarez", "diego costa", "alexandre lacazette"))) + ggtitle('Most P
roductive Shooters in Europe') + xlab('Numbers') + ylab('Shooters') + theme(legend.title =
element_blank()) + scale_fill_discrete(labels = c("goal", "shoot"))
```



Who scored the most goals in each league?

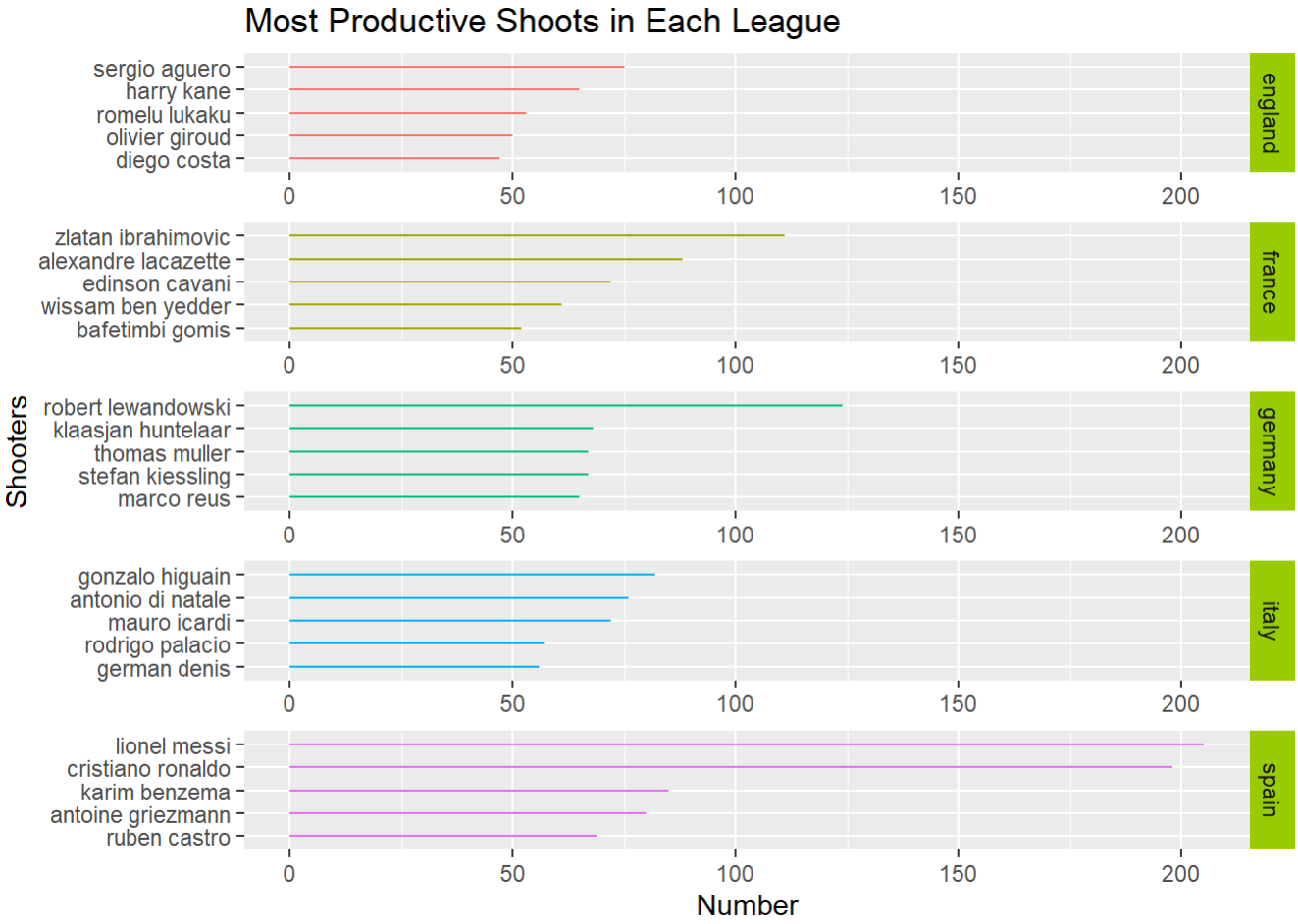
The second topic coming into my mind is to explore the most productive shooters in each league. After a series of data cleaning, a plot reflecting the most productive shooters in each league is created. It seems that there are no 'super shooters' in England and France, the variance of number of goals is smaller. Attackers in Spain seems scored more goals than others.

```
# b is a temporary data set containing the most productive shoots in each league
f <- function(x){arrange(x)[1:5,]}
a <- filter(events, is_goal == 1) %>% group_by(player, country) %>% summarise(number = n())
 %>% arrange(desc(number))
```

```
## `summarise()` has grouped output by 'player'. You can override using the `.groups` argum
ent.
```

```
b <- by(a, a$country, f)
```

```
# shot.data is the tibble we use to get the second plot
shot.data <- NULL
for (i in 1:5){
  temp <- by(a, a$country, f)[i][[1]]
  shot.data <- rbind(shot.data, temp)
}
shot.data$player <- factor(shot.data$player, levels = shot.data$player[order(shot.data$numb
er)])

# plot 2, "Most Productive Shooters in Each League"
ggplot(shot.data, aes(x=player, y=number, color = country))+ geom_segment(aes(x=player,xend
=player, y=0, yend=number)) + facet_wrap(~country, scales = "free", nrow = 5, strip.positio
n = 'right') + ylim(0, 205) + coord_flip() + theme(strip.background = element_rect(fill='#9
9CC00'), legend.position = "none") + ylab('Number') + xlab('Shooters') + ggtitle("Most Prod
uctive Shoots in Each League")
```



Comparison among different leagues

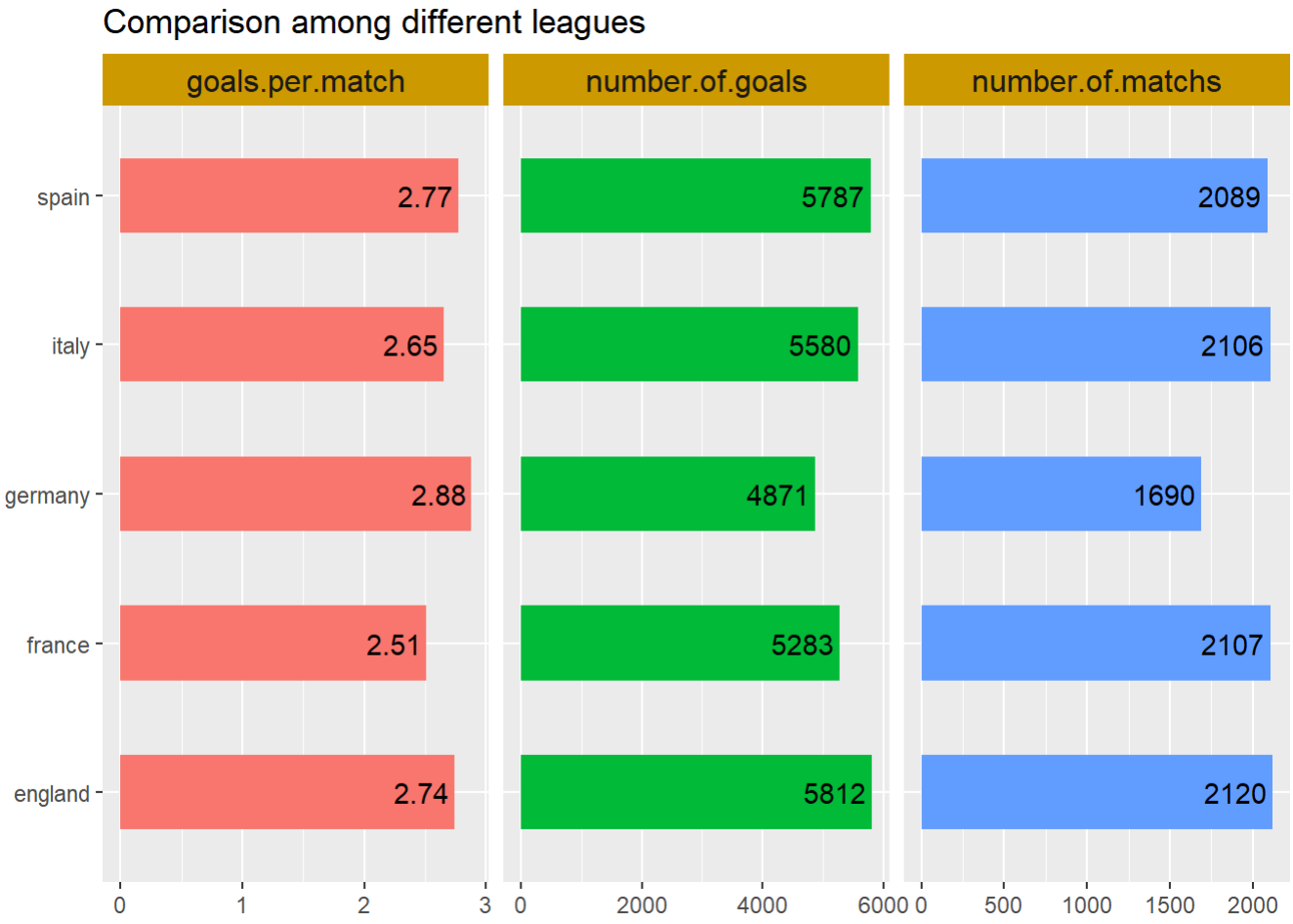
The different performance of attack players in different leagues leads me to think about this question: is there any significant difference among this leagues? For example, it might be that the football style in Spain is more emphasis on offensive. That's why Lionel Messi and Cristiano Ronaldo could score much more goals than others. Or the football philosophy in England is defend, so players in England could not get as many as scores as other leagues.

To clarify this question, I compared the average goals per match of each league which showed that there were no significant difference. Therefore, it might be owning to the diversity of attack method, the goals were more equally distributed to players in England. Besides, Average goals per match in Spain were not really high, which means Lionel Messi and Cristiano Ronaldo are indeed excellent players.

```
# goals.data contains the number of matches, goals and the goals per match in each league
goals.data <- ginf %>% group_by(country) %>% summarise(number.of.matchs = n())
ginf$goal <- ginf$fthg + ginf$ftag
goals.data <- goals.data %>% left_join(ginf %>% group_by(country) %>% summarise(number.of.goals = sum(goal)))
```

```
## Joining, by = "country"
```

```
goals.data$goals.per.match <- goals.data$number.of.goals/goals.data$number.of.matchs
goals.data$goals.per.match <- round(goals.data$goals.per.match, digits = 2)
goals.data <- pivot_longer(goals.data, 2:4, names_to = 'type', values_to = 'number')

# plot 3
ggplot(goals.data, aes(country, number, fill = type)) + geom_bar(stat="identity",position="dodge", width=.5) + coord_flip() + facet_wrap(~type, ncol = 3, scales = "free_x") + geom_text(aes(label = number), hjust = 1.1) + theme(strip.background = element_rect(fill='#CC9900'), strip.text = element_text(size = 12), legend.position = "none") + ggtitle('Comparison among different leagues') + xlab(NULL) + ylab(NULL)
```



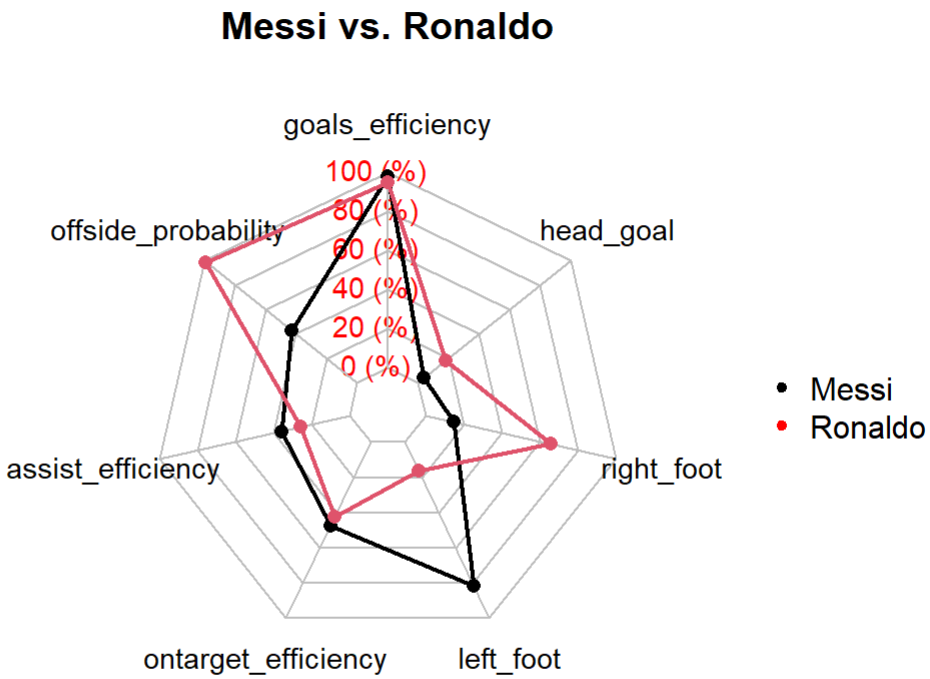
Lionel Messi vs. Cristiano Ronaldo

Being the most capable contemporary football players, Lionel Messi and Cristiano Ronaldo achieved unprecedented successful. They are both excellent Offensive players, but is there any difference between them? That's what I want to answer in this plot.

Radar chart is an ideal method to make comparison. We can see from this chart that they both achieve quite high goal efficiency(number of goals/number of attempts) and on target efficiency(number of on target shoots/number of attempts). Cristiano Ronaldo tends to be a pure shooter. His offside frequency is much higher than Messi while Messi

tends to be an organizer. That's why Messi's assist number is higher than Ronaldo. Moreover, the form of scoring of Ronaldo is more diverse while most goals of Messi are contributed by his left foot. Perhaps it's the most expensive left foot in the world!

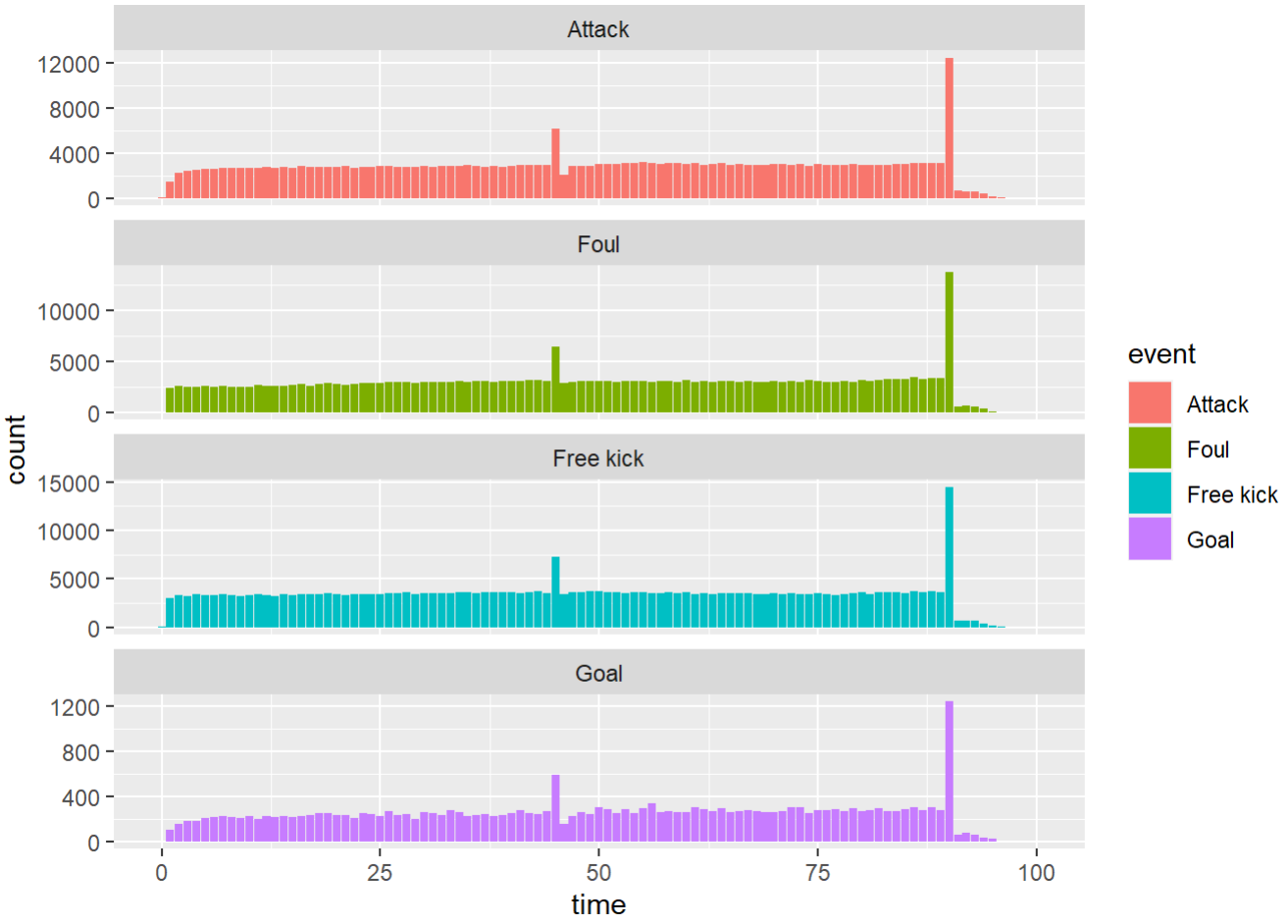
```
messi.Ronaldo <- tibble(name = c('Messi', 'Ronaldo'), goals = NA, matches = NA, left = NA,
right = NA, head = NA, offside = NA, ontarget = NA, assist = NA)
messi.Ronaldo$goals <- c(nrow(filter(events, player == 'lionel messi' & is_goal == 1)), nro
w(filter(events, player == 'cristiano ronaldo' & is_goal == 1)))
messi.Ronaldo$matches <- c((filter(ginf, at == 'Barcelona') %>% summarise(count = n()) + fi
lter(ginf, ht == 'Barcelona') %>% summarise(count = n()))[1,1],  (filter(ginf, at == 'Real
Madrid') %>% summarise(count = n()) + filter(ginf, ht == 'Real Madrid') %>% summarise(count
 = n()))[1,1])
messi.Ronaldo$left <- c(nrow(filter(events, player == 'lionel messi' & is_goal == 1 & bodyp
art == 2)), nrow(filter(events, player == 'cristiano ronaldo' & is_goal == 1 & bodypart ==
2)))
messi.Ronaldo$right <- c(nrow(filter(events, player == 'lionel messi' & is_goal == 1 & body
part == 1)), nrow(filter(events, player == 'cristiano ronaldo' & is_goal == 1 & bodypart ==
 1)))
messi.Ronaldo$head <- c(nrow(filter(events, player == 'lionel messi' & is_goal == 1 & bodyp
art == 3)), nrow(filter(events, player == 'cristiano ronaldo' & is_goal == 1 & bodypart ==
3)))
messi.Ronaldo$offside <- c(nrow(filter(events, player == 'lionel messi' & event_type == 9))
, nrow(filter(events, player == 'cristiano ronaldo' & event_type == 9)))
messi.Ronaldo$ontarget <- c(nrow(filter(events, player == 'lionel messi' & shot_outcome ==
1)), nrow(filter(events, player == 'cristiano ronaldo' &  shot_outcome == 1)))
messi.Ronaldo$assist <- c(nrow(filter(events, player2 == 'lionel messi' & is_goal == 1)), n
row(filter(events, player2 == 'cristiano ronaldo' & is_goal == 1)))
messi.Ronaldo$shot <- c(nrow(filter(events, player == 'lionel messi' & event_type == 1)), n
row(filter(events, player == 'cristiano ronaldo' & event_type == 1)))

messi.Ronaldo$goals_efficiency <- messi.Ronaldo$goals/messi.Ronaldo$matches
messi.Ronaldo$offside_probability <- messi.Ronaldo$offside/messi.Ronaldo$matches
messi.Ronaldo$assist_efficiency <- messi.Ronaldo$assist/messi.Ronaldo$matches
messi.Ronaldo$ontarget_efficiency <- messi.Ronaldo$ontarget/messi.Ronaldo$shot
messi.Ronaldo$left_foot <- messi.Ronaldo$left/messi.Ronaldo$goals
messi.Ronaldo$right_foot <- messi.Ronaldo$right/messi.Ronaldo$goals
messi.Ronaldo$head_goal <- messi.Ronaldo$head/messi.Ronaldo$goals

dat <- matrix(c(rep(1, 7), rep(0, 7)), nrow = 2, byrow = T)
colnames(dat) <- c('goals_efficiency', 'offside_probability', 'assist_efficiency', 'ontarge
t_efficiency', 'left_foot', 'right_foot', 'head_goal')

dat <- as.data.frame(rbind(dat, messi.Ronaldo[, 11:17]))

radarchart(dat, axistype =1,seg=5,pty=16,plty=1,plwd = 2, cglty = 1,cglcol = "grey",axislab
col = "red",title = "Messi vs. Ronaldo", vlcex = 0.9,calcex = 0.9,palcex = 0.5)
legend(x = "right", legend = c('Messi', 'Ronaldo'), seg.len=0.5, bty = "n", horiz=FALSE, pc
h = 20 , col = c("black", "red"))
```



Which part in one match is can not be missed?

One football match will roughly last 2 hours, which might be too long to busy people. One strategy is that people can only watch the most exciting part. The data reveals that the last part of each half is the most intriguing. Events are concentrated in the final stage of the game, no matter what kind of events: goals, fouls and attempts. Therefore, as a football fan, no matter how busy you are, please do not miss the last part of a game.

```
dat <- filter(events, event_type %in% c(1, 9)) %>% dplyr::select(time, event_type)
dat$event <- rep('Attack', nrow(dat))

dat1 <- filter(events, event_type %in% c(2, 8, 11)) %>% dplyr::select(time, event_type)
dat1$event <- rep('Free kick', nrow(dat1))

dat2 <- filter(events, event_type %in% c(3, 4, 5, 6, 10)) %>% dplyr::select(time, event_typ
e)
dat2$event <- rep('Foul', nrow(dat2))

dat <- bind_rows(dat, dat1, dat2)

dat3 <- filter(events, is_goal == 1) %>% dplyr::select(time, is_goal)
dat3$event <- rep('Goal', nrow(dat3))

dat <- rbind(dplyr::select(dat, time, event), dplyr::select(dat3, time, event))

ggplot(data = dat, aes(time, fill = event)) + geom_bar() + facet_wrap(~event, nrow = 4, sca
les = "free_y")
```



Use statistical model to predict team's ability

After a series of data visualization, I want to predict team's ability by the events. To construct the model, I make up some derivative index:

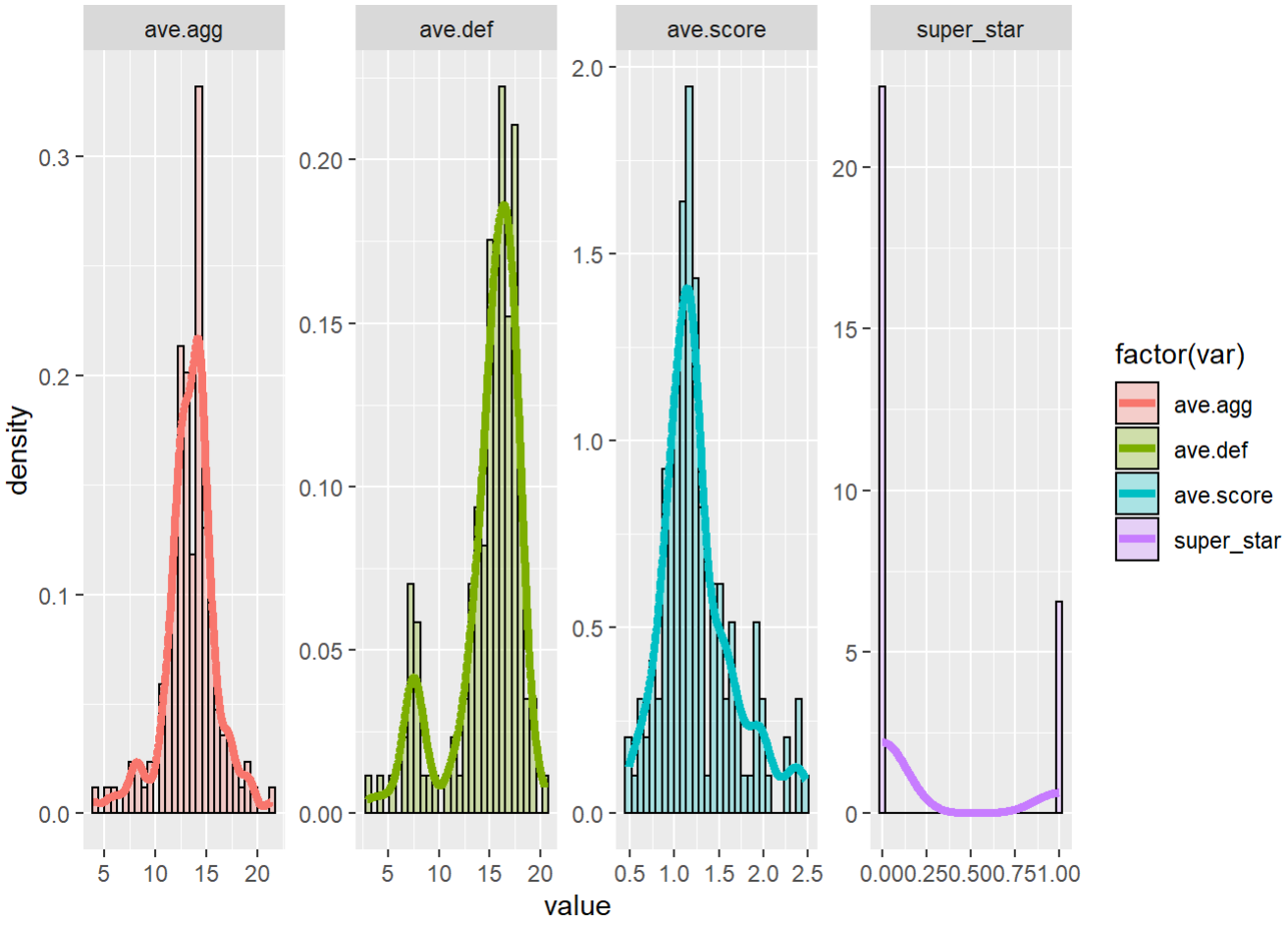
1. Average Score(response): 3 scores for win, 1 score for draw and 0 for lose. average Score is an objective description of one team's ability.
2. Offensive. performance: it's an index of attack events, like goals, key pass and failed key pass. Higher this index, more offensive.
3. Defensive strength: it's an index containing the number of foul, yellow card, red card and hand ball. Higher this index, the way of playing soccer is more aggressive.
4. Super star: it's a dummy variable which will be 1 for those teams having super shooter and will be 0 for those teams doesn't. The definition of super shooter is one player who is listed in the top 5 most goals table.

I run linear model to fit average Score to offensive, defensive strength and super star. The adjusted R-squared is 0.6641 which is very meaningful. However, I need to mention that there are a lot of pitfalls in this regression. For example, the super star variable. The regression shows that the coefficient is statistical significant, but it might be because the teams themselves are quite strong,

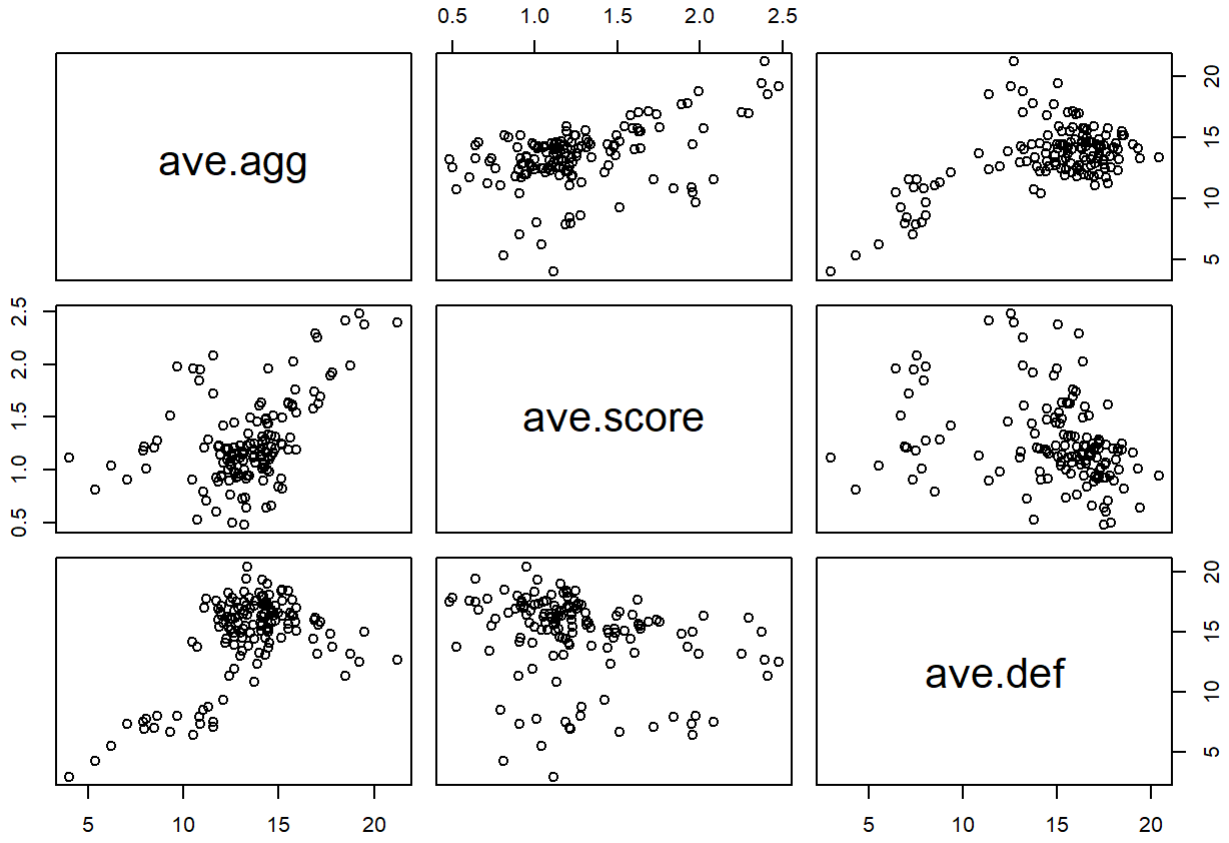
```
# create a tibble 'score' to summarize total score of each team
ht.score <- select(ginf, ht, htscore)
at.score <- select(ginf, at, atscore)
colnames(ht.score) <- c('team', 'score')
colnames(at.score) <- c('team', 'score')
score <- bind_rows(ht.score, at.score) %>% group_by(team) %>% summarise(sum.score = sum(sco
re))
```

```
temp <- select(ginf, id_odsp, ht, at)
score <- score %>% left_join(pivot_longer(temp, 2:3) %>% group_by(value) %>% summarise(coun
t = n()), by = c('team' = 'value'))

# attack performance
score <- score %>% left_join(filter(events, event_type == 1 | event_type2 == 12 | event_typ
e2 == 13) %>% group_by(event_team) %>% summarise(aggressive = n()), by = c('team' = 'event_
team'))

# Defensive strength
score <- score %>% left_join(filter(events, event_type %in% c(3, 4, 5, 6, 10)) %>% group_by
(event_team) %>% summarise(def = n()), by = c('team' = 'event_team'))

# finger the league each team belongs to
team <- select(ginf, country, ht, at) %>% pivot_longer(2:3, names_to = 'type', values_to =
'teams') %>% distinct(teams, .keep_all = T)
team <- select(team, -'type')
score <- left_join(score, team, by = c('team' = 'teams'))

# whether a team has super shooter
super.shooter <- events %>% filter(is_goal == 1 & event_type2 != 15) %>% select(event_team,
 player) %>% distinct(player, event_team, .keep_all = T) %>% inner_join(shot.data, by = 'pl
ayer') %>% select(event_team, player)
score$super_star <- 0
for (i in 1:nrow(score)){
  if (score$team[i] %in% super.shooter$event_team){score$super_star[i] <- 1}
}

# omit NA
score <- setdiff(score, score[complete.cases(score) == 0, ])

# calculate average index
score$ave.score <- score$sum.score/score$count
score$ave.agg <- score$aggressive/score$count
score$ave.def <- score$def/score$count

# plot the density of covariates
score %>% gather(ave.agg, ave.score, ave.def, super_star, key = "var", value = "value") %>%
 ggplot(aes(x = value)) + geom_histogram(aes(fill = factor(var), y = ..density..), alpha =
0.3, colour = 'black') + stat_density(geom = 'line', position = 'identity', size = 1.5, aes
(colour = factor(var))) +  facet_wrap(~ var, scales = "free", ncol = 4)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# plot the correlation among covariates
pairs(~ave.agg+ave.score+ave.def,data = score)
```



```
# linear model
score <- score %>% mutate(ave.score = sum.score/count, attack_ability = aggressive/count, d
efensive_strength = def/count)
lm.fit <- lm(ave.score ~ ave.agg + ave.def + super_star, data = score)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = ave.score ~ ave.agg + ave.def + super_star, data = score)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58720 -0.11067  0.00946  0.13763  0.79166
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.718323   0.111635   6.435 1.89e-09 ***
## ave.agg      0.115409   0.011026  10.467  < 2e-16 ***
## ave.def     -0.072391   0.007891  -9.174 5.87e-16 ***
## super_star   0.248887   0.057169   4.354 2.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2371 on 138 degrees of freedom
## Multiple R-squared:  0.6641, Adjusted R-squared:  0.6568
## F-statistic: 90.94 on 3 and 138 DF,  p-value: < 2.2e-16
```

Model diagnostics

To detect the performance of this model, I run 500 times bootstrap to create CI for all predictors and the result is quite similar to the report of linear model. Cross-validation shows that the MSE of this model is 0.238.

```
# bootstrap to create CI
set.seed(1234)
boot <- bootstraps(score, 500)
boot_result <- map(boot$splits, ~tidy(lm(ave.score ~ ave.agg + ave.def + super_star, data =
 .))) %>% bind_rows(.)
boot_CI <- boot_result %>% group_by(term) %>% summarize(conf.low = quantile(estimate, 0.05
/ 2), conf.high = quantile(estimate, 1 - 0.05 / 2))
boot_CI
```

```
## # A tibble: 4 x 3
##   term        conf.low conf.high
## * <chr>          <dbl>     <dbl>
## 1 (Intercept)   0.502     0.932
## 2 ave.agg       0.0878    0.139
## 3 ave.def      -0.0884   -0.0558
## 4 super_star    0.130     0.381
```

```
# Cross Validation of the model
score.cv <- crossv_kfold(score, k = 10) %>% mutate(model = map(train, ~lm(ave.score ~ ave.a
gg + ave.def + super_star, data = .)))
map2_dbl(score.cv$model, score.cv$test,rmse) %>% mean()
```

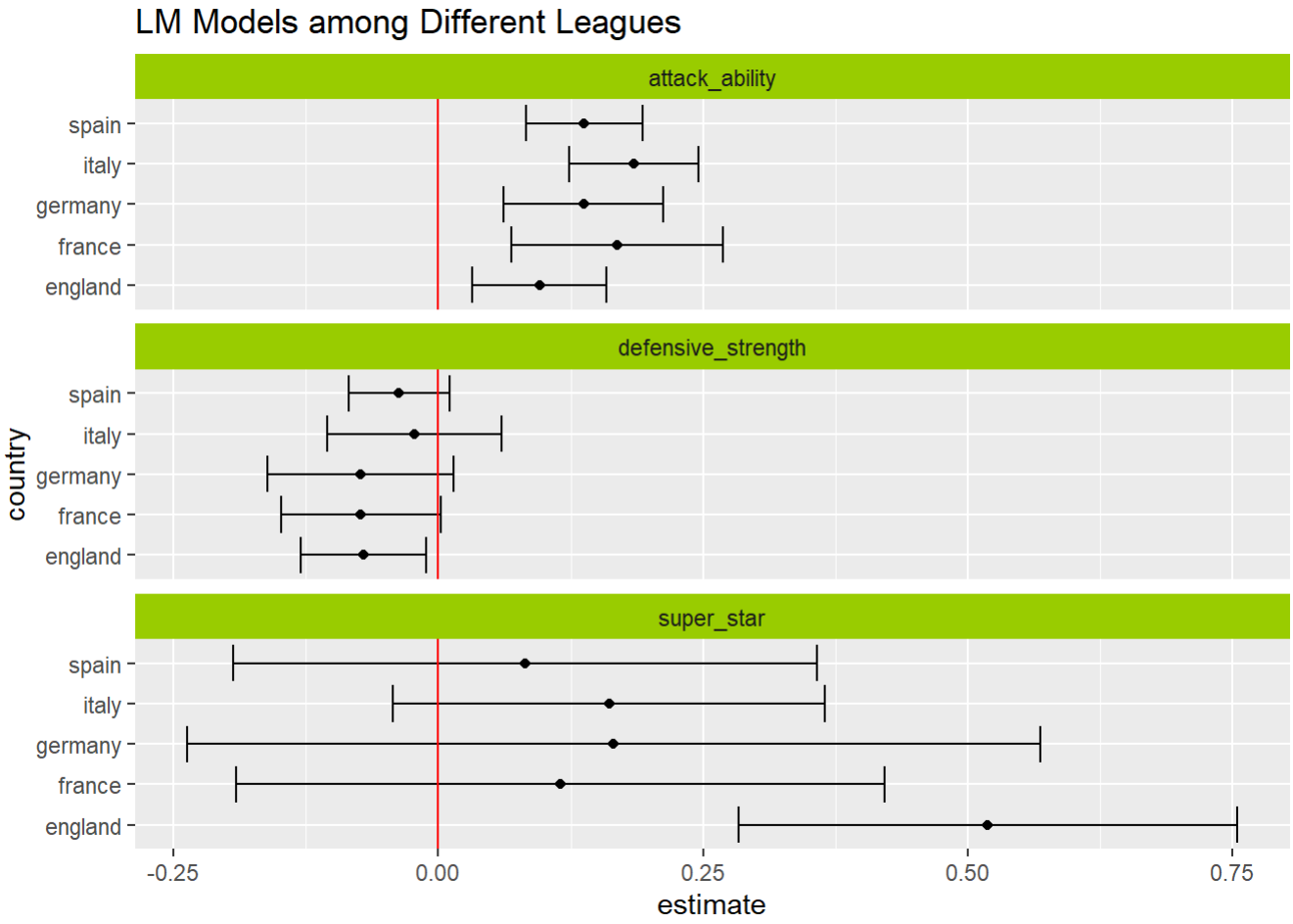
```
## [1] 0.2379944
```

Fit model to each league

In this chapter, I want to explore that how this regression model performs to different leagues. The data shows that there are no significant difference among the performance. All the residuals of this models are symmetric.

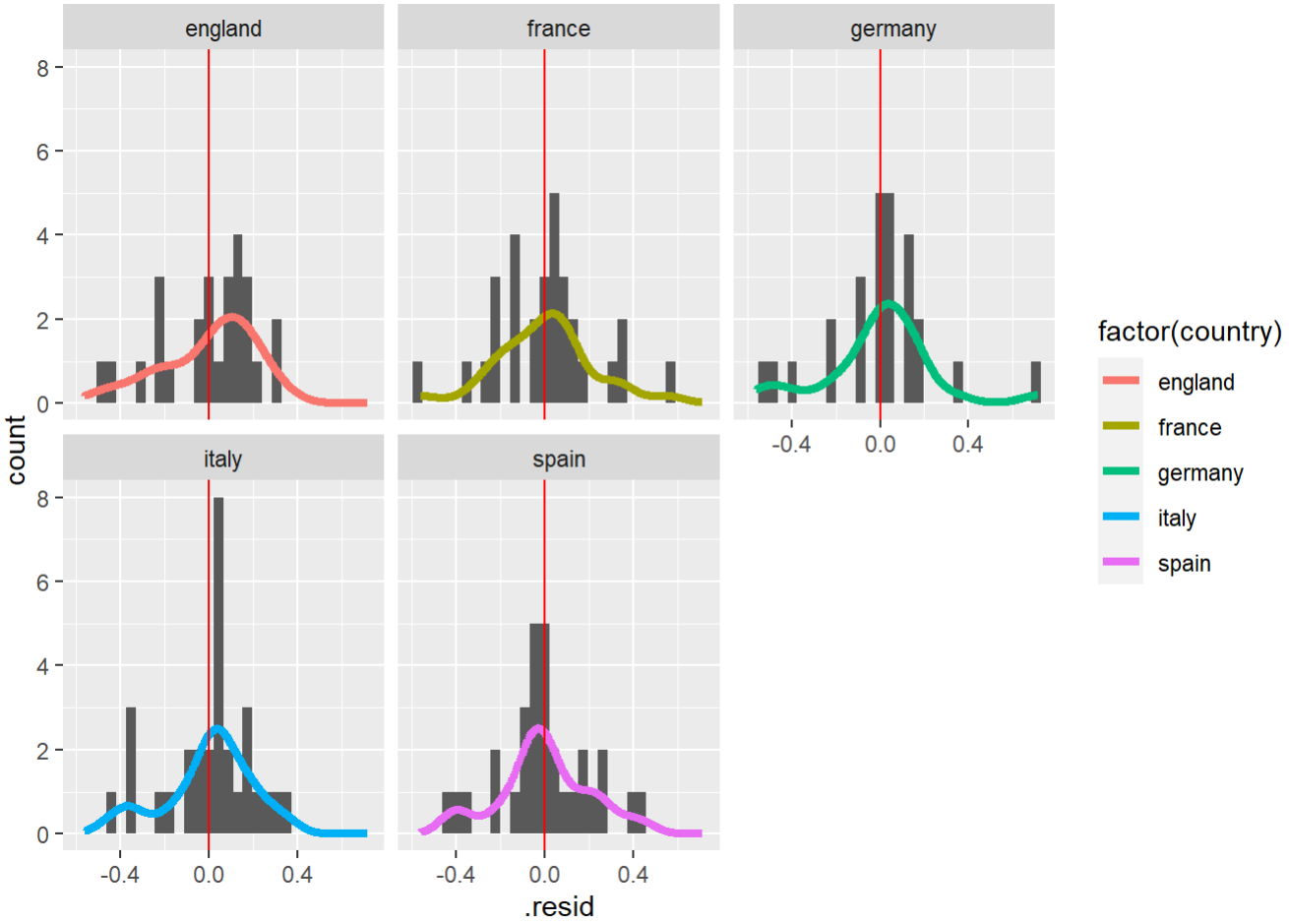
This regression model seems a little bit better in England, Italy and Spain, since the R-squared are higher and the confidence intervals are smaller.

```
# lm models to each league
regressions <- score %>% group_by(country) %>% do(tidy(lm(ave.score ~ attack_ability + defe
nsive_strength + super_star, data = .), conf.int = TRUE))
coefs <- regressions %>% ungroup()
coefs %>% filter(term != '(Intercept)') %>% ggplot(aes(x = estimate, y = country)) + geom_p
oint() + geom_errorbarh(aes(xmin = conf.low, xmax = conf.high)) + facet_wrap(~ term, nrow =
 3) + geom_vline(xintercept = 0, color = "red") + theme(strip.background = element_rect(fil
l='#99CC00')) + ggtitle('LM Models among Different Leagues')
```




```
# residuals of each team according to different leagues
reg_observations <- score %>% group_by(country) %>% do(augment(lm(ave.score ~ attack_abilit
y + defensive_strength + super_star, data = .))) %>% left_join(dplyr::select(score, team, a
ve.score, ave.agg), by = c('ave.score' = 'ave.score', 'attack_ability' = 'ave.agg'))
dplyr::select(reg_observations, .resid, team, country) %>% ggplot() + geom_histogram(aes(.r
esid)) + stat_density(geom = 'line', position = 'identity', size = 1.5, aes(.resid, color =
 factor(country))) + facet_wrap(~country) + geom_vline(xintercept = 0, color = "red")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# R-squared
R2 <- score %>% group_by(country) %>% do(glance(lm(ave.score ~ attack_ability + defensive_s
trength + super_star, data = .)))
R2[, 1:2]
```

```
## # A tibble: 5 x 2
## # Groups:   country [5]
##   country r.squared
##   <chr>       <dbl>
## 1 england     0.714
## 2 france      0.599
## 3 germany     0.698
## 4 italy       0.783
## 5 spain       0.752
```