2021/5/5

The Beauty of Soccer

Yining Wang

About the data set

- A data set containing over 900,000 events such as shoots, goals, fouls, and penalties from 9,074 football games across 5 most influential Europe leagues will be used in this project. The link is showed below:

- https://www.kaggle.com/ahmedyoussef/the-beautiful-game-analysis-of-football-events

Data frame 'events'

```
> glimpse(events)
Rows: 941,009
Columns: 22
$ id_odsp       <chr> "UFot0hit/", "UFot0hit/", "UFot0hit/", "UFot0hit/", "UFo...
$ id_event      <chr> "UFot0hit1", "UFot0hit2", "UFot0hit3", "UFot0hit4", "UFo...
$ sort_order    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
$ time          <dbl> 2, 4, 4, 7, 7, 9, 10, 11, 11, 13, 14, 14, 14, 17, 19, 20...
$ text          <chr> "Attempt missed. Mladen Petric (Hamburg) left footed sho...
$ event_type    <dbl> 1, 2, 2, 3, 8, 10, 2, 8, 3, 3, 8, 1, 3, 1, 1, 3, 8, 1, 1...
$ event_type2   <dbl> 12, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 12, NA, 12, ...
$ side          <dbl> 2, 1, 1, 1, 2, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 1, 2, 2, 1,...
$ event_team    <chr> "Hamburg SV", "Borussia Dortmund", "Borussia Dortmund", ...
$ opponent      <chr> "Borussia Dortmund", "Hamburg SV", "Hamburg SV", "Hambur...
$ player        <chr> "mladen petric", "dennis diekmeier", "heiko westermann",...
$ player2       <chr> "gokhan tore", "dennis diekmeier", "heiko westermann", N...
$ player_in     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
$ player_out    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
$ shot_place    <dbl> 6, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 13, NA, 4, 2,...
$ shot_outcome  <dbl> 2, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 2, NA, 1, 3, ...
$ is_goal       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,...
$ location      <dbl> 9, NA, NA, NA, 2, NA, NA, 2, NA, NA, 4, 15, NA, 9, 15, N...
$ bodypart      <dbl> 2, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 1, NA, 2, 1, ...
$ assist_method <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1,...
$ situation     <dbl> 1, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 1, NA, 1, 1, ...
$ fast_break    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

Dictionary

event_type

0	Announcement
1	Attempt
2	Corner
3	Foul
4	Yellow card
5	Second yellow card
6	Red card
7	Substitution
8	Free kick won
9	Offside
10	Hand ball
11	Penalty conceded

event_type2

12	Key Pass
13	Failed through ball
14	Sending off
15	Own goal

side

1	Home
2	Away

shot_place

1	Bit too high
2	Blocked
3	Bottom left corner
4	Bottom right corner
5	Centre of the goal
6	High and wide
7	Hits the bar
8	Misses to the left
9	Misses to the right
10	Too high
11	Top centre of the goal
12	Top left corner
13	Top right corner

shot_outcome

1	On target
2	Off target
3	Blocked
4	Hit the bar

assist_method

0	None
1	Pass
2	Cross
3	Headed pass
4	Through ball

situation

1	Open play
2	Set piece
3	Corner
4	Free kick

location

1	Attacking half
2	Defensive half
3	Centre of the box
4	Left wing
5	Right wing
6	Difficult angle and long range
7	Difficult angle on the left
8	Difficult angle on the right
9	Left side of the box
10	Left side of the six yard box
11	Right side of the box
12	Right side of the six yard box
13	Very close range
14	Penalty spot
15	Outside the box
16	Long range
17	More than 35 yards
18	More than 40 yards
19	Not recorded

bodypart

1	right foot
2	left foot
3	head

Data frame 'ginf'

```
> glimpse(ginf)
Rows: 10,112
Columns: 18
$ id_odsp    <chr> "UFot0hit/", "Aw5DflLH/", "bkjpaC6n/", "CzPV312a/", "GUOdmtI...
$ link_odsp  <chr> "/soccer/germany/bundesliga-2011-2012/dortmund-hamburger-UFo...
$ adv_stats  <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, ...
$ date       <date> 2011-08-05, 2011-08-06, 2011-08-06, 2011-08-06, 2011-08-06,...
$ league     <chr> "D1", "D1", "D1", "F1", "F1", "D1", "F1", "F1", "F1", "D1", ...
$ season     <dbl> 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, ...
$ country    <chr> "germany", "germany", "germany", "france", "france", "german...
$ ht         <chr> "Borussia Dortmund", "FC Augsburg", "Werder Bremen", "Paris ...
$ at         <chr> "Hamburg SV", "SC Freiburg", "Kaiserslautern", "Lorient", "V...
$ fthg       <dbl> 3, 2, 2, 0, 1, 0, 2, 0, 1, 0, 1, 3, 3, 2, 2, 1, 1, 2, 0, 0, ...
$ ftag       <dbl> 1, 2, 0, 1, 0, 1, 2, 2, 3, 3, 1, 1, 0, 1, 2, 2, 5, 0, 1, 0, ...
$ odd_h      <dbl> 1.56, 2.36, 1.83, 1.55, 2.50, 2.06, 2.29, 2.80, 4.50, 3.00, ...
$ odd_d      <dbl> 4.41, 3.60, 4.20, 4.50, 3.40, 3.75, 3.25, 3.10, 3.55, 3.80, ...
$ odd_a      <dbl> 7.42, 3.40, 4.80, 9.40, 3.45, 3.95, 3.85, 3.05, 2.00, 2.54, ...
$ odd_over   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
$ odd_under  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
$ odd_bts    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
$ odd_bts_n  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
```
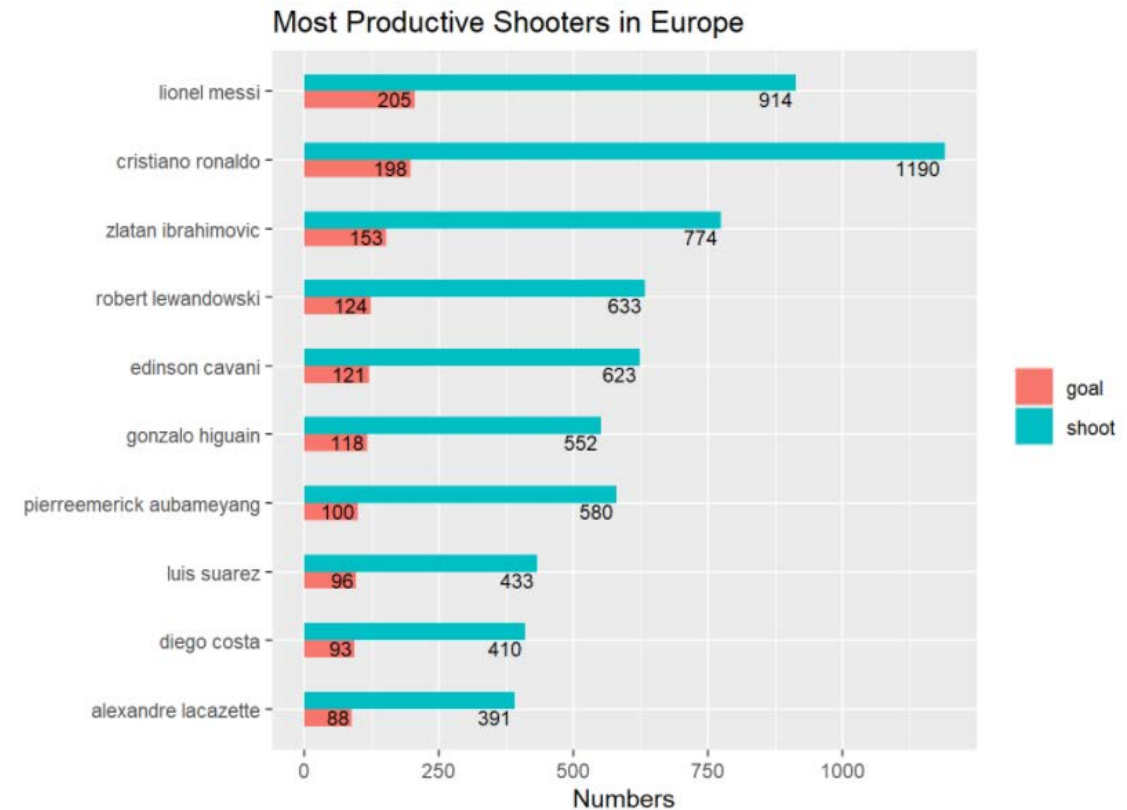
Who scored the most goals?

	player	ontarget.number	shot.number	goal.number	on.target.percentage	goal.percentage
1	miroslav klose	91	205	51	0.444	0.249
2	carlos bacca	113	230	57	0.491	0.248
3	diego costa	191	410	93	0.466	0.227
4	alexandre lacazette	193	391	88	0.494	0.225
5	lionel messi	435	914	205	0.476	0.224
6	mauro icardi	122	323	72	0.378	0.223
7	luis suarez	201	433	96	0.464	0.222
8	mario gomez	98	221	48	0.443	0.217
9	gonzalo higuain	261	552	118	0.473	0.214
10	falcao	171	381	80	0.449	0.210

Most Productive Shooters in Europe

2021/5/5

Who scored the most goals in each league?

- There are no 'super shooters' in England and France

- Players in Spain scored much more goals than others

- Is there any difference in football philosophy among this leagues? Maybe people tend to focus on offense but not on defense in Spain?


Most Productive Shoots in Each League
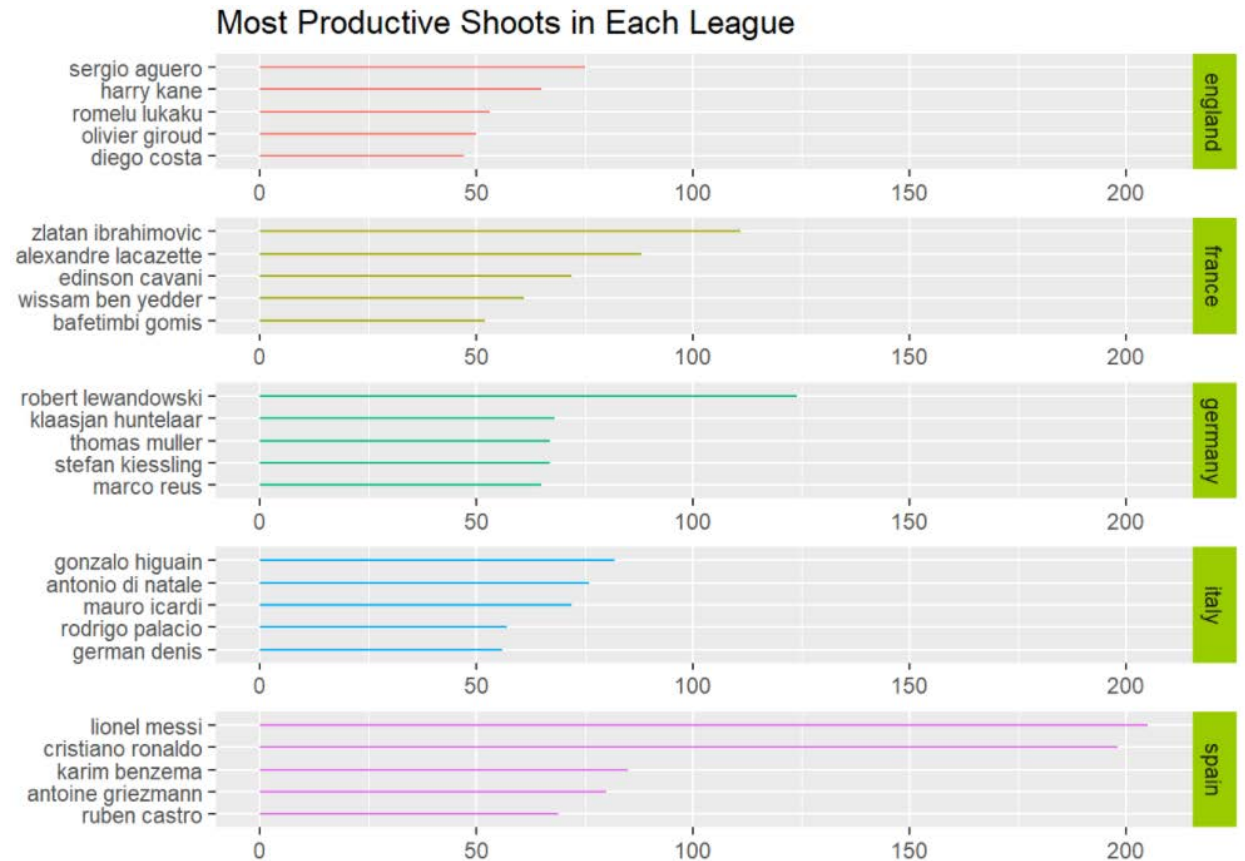
Comparison among different leagues

- There are no significant difference in the number of goals per match

- The number of goals per match is above average in England, therefore, the goals were more equally distributed to players there

- Average goals per match in Spain were not really high

Lionel Messi vs. Cristiano Ronaldo

- They both achieve quite high goal efficiency and on target efficiency

- Cristiano Ronaldo tends to be a pure shooter. His offside frequency is much.

- Messi tends to be an organizer. His assist number is higher than Ronaldo.

- The form of scoring of Ronaldo is more diverse while most goals of Messi are contributed by his left foot.


Messi vs. Ronaldo
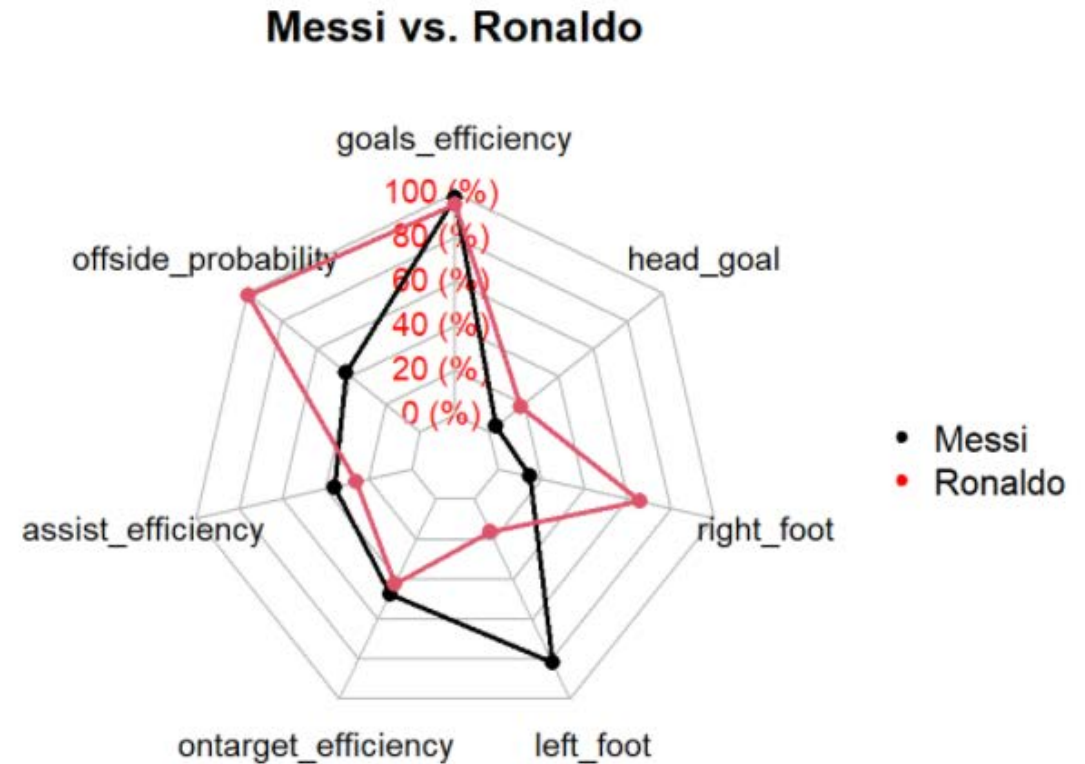
No matter how busy you are, DO NOT miss the last part!
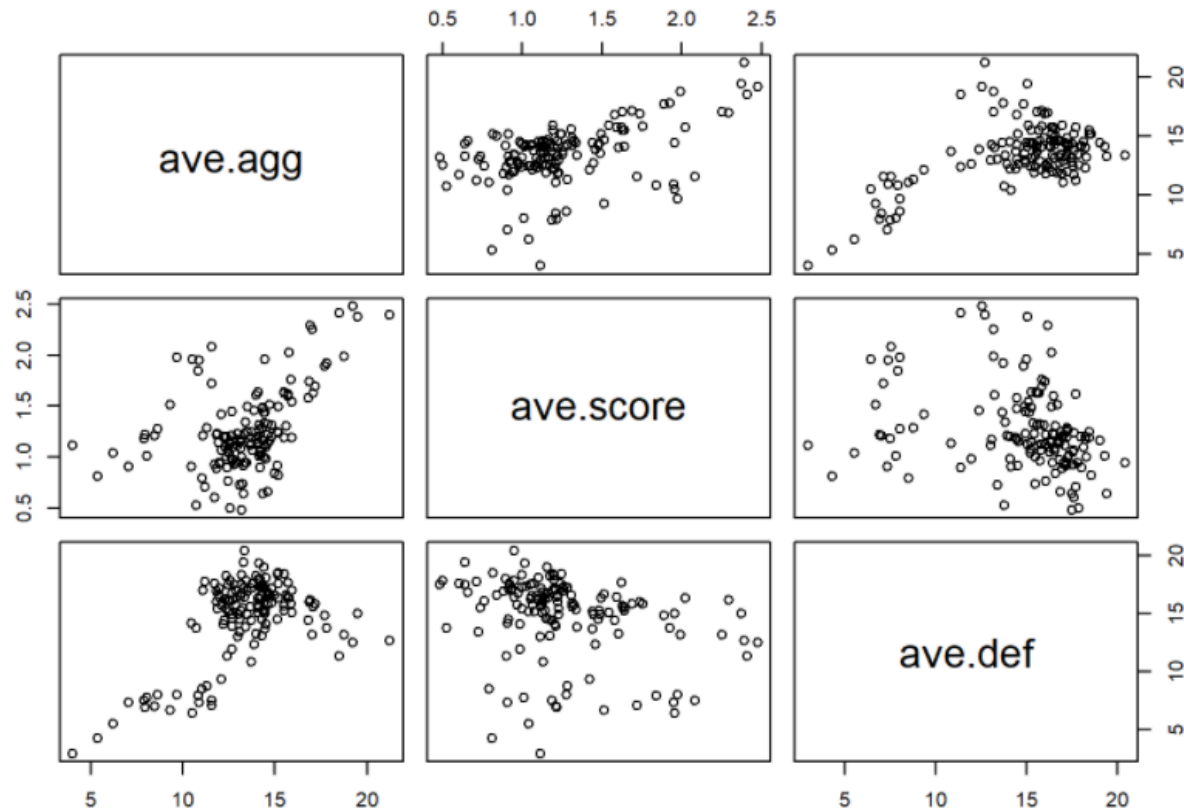

2021/5/5
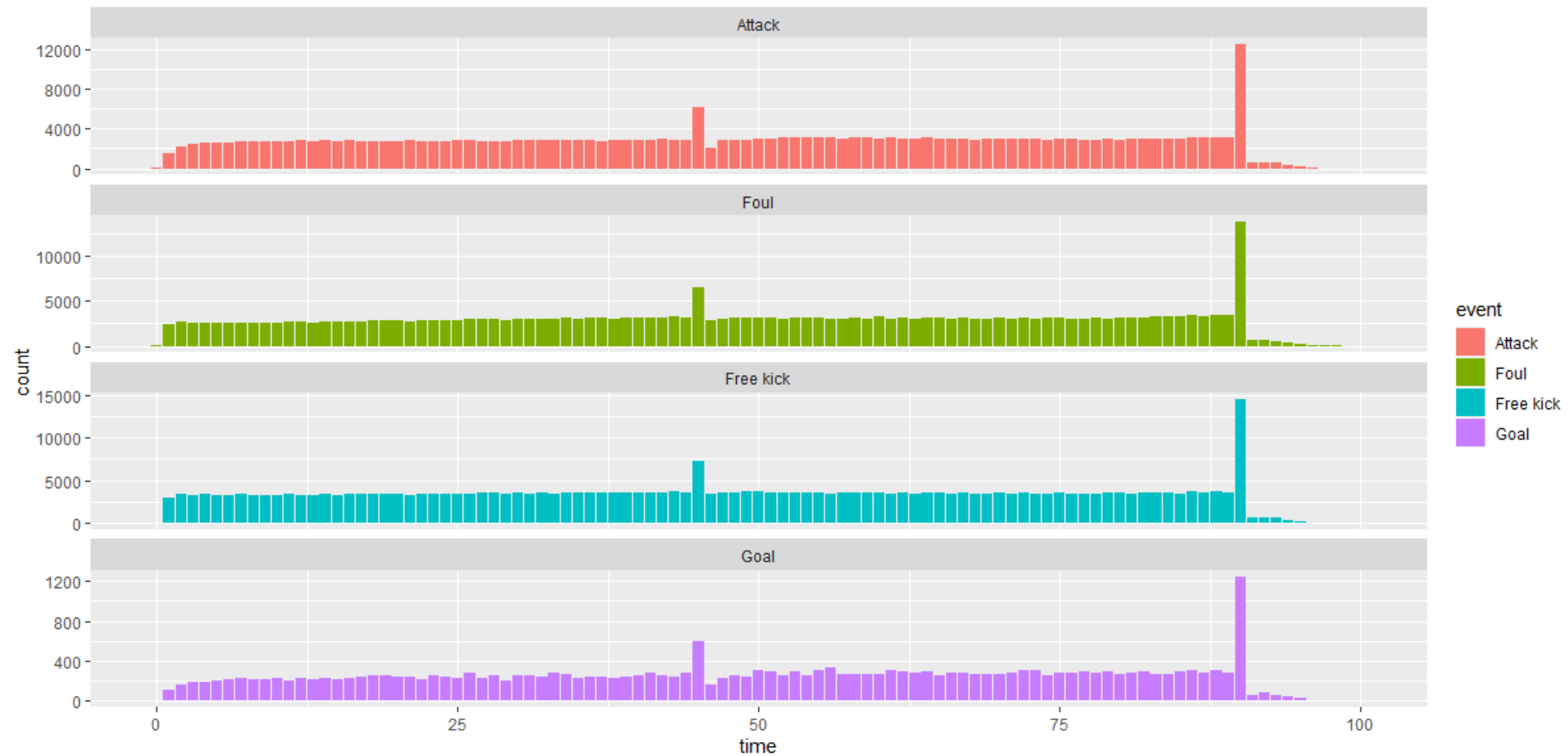
Predict team's performance

Make up index

- Average Score(response): 3 scores for win, 1 score for draw and 0 for lose.

- Offensive performance: goals, key pass and failed key pass.

- Defensive strength: foul, yellow card, red card and hand ball.

- Super star: 1 for those teams having super shooter and will be 0 for those teams don't.
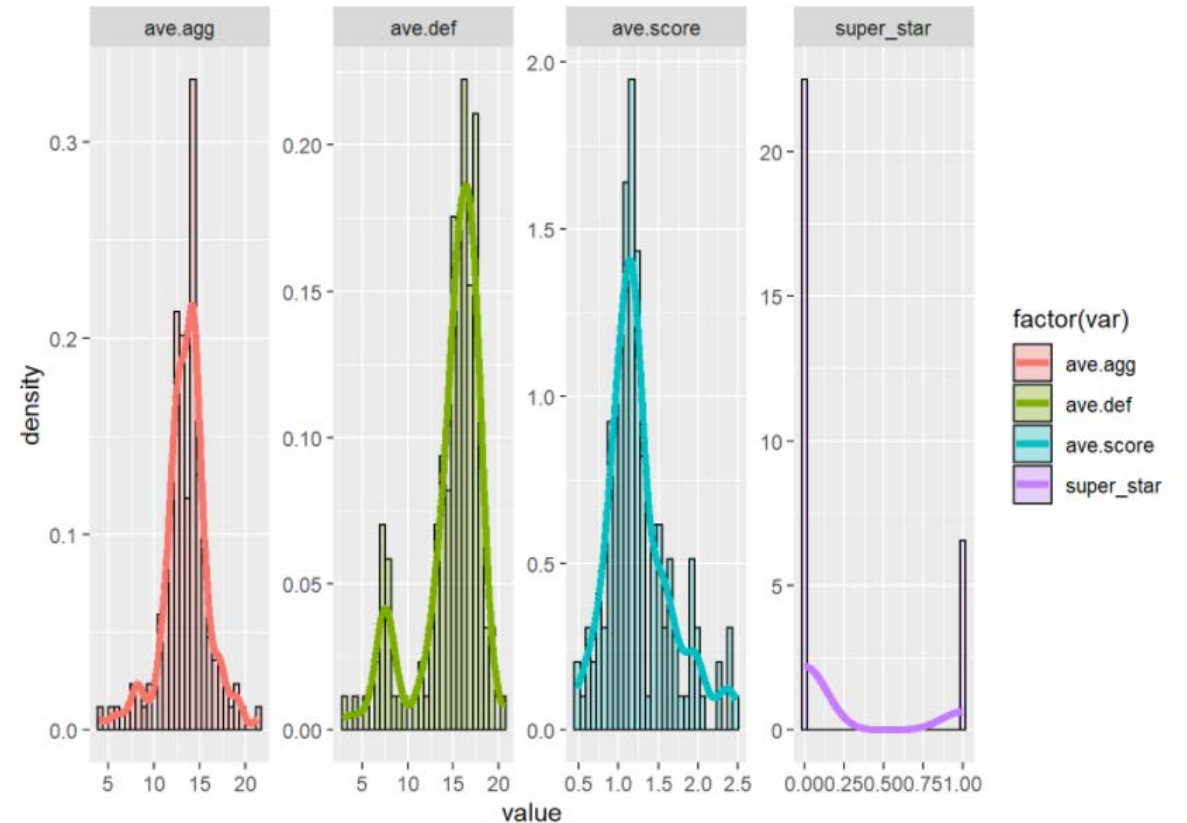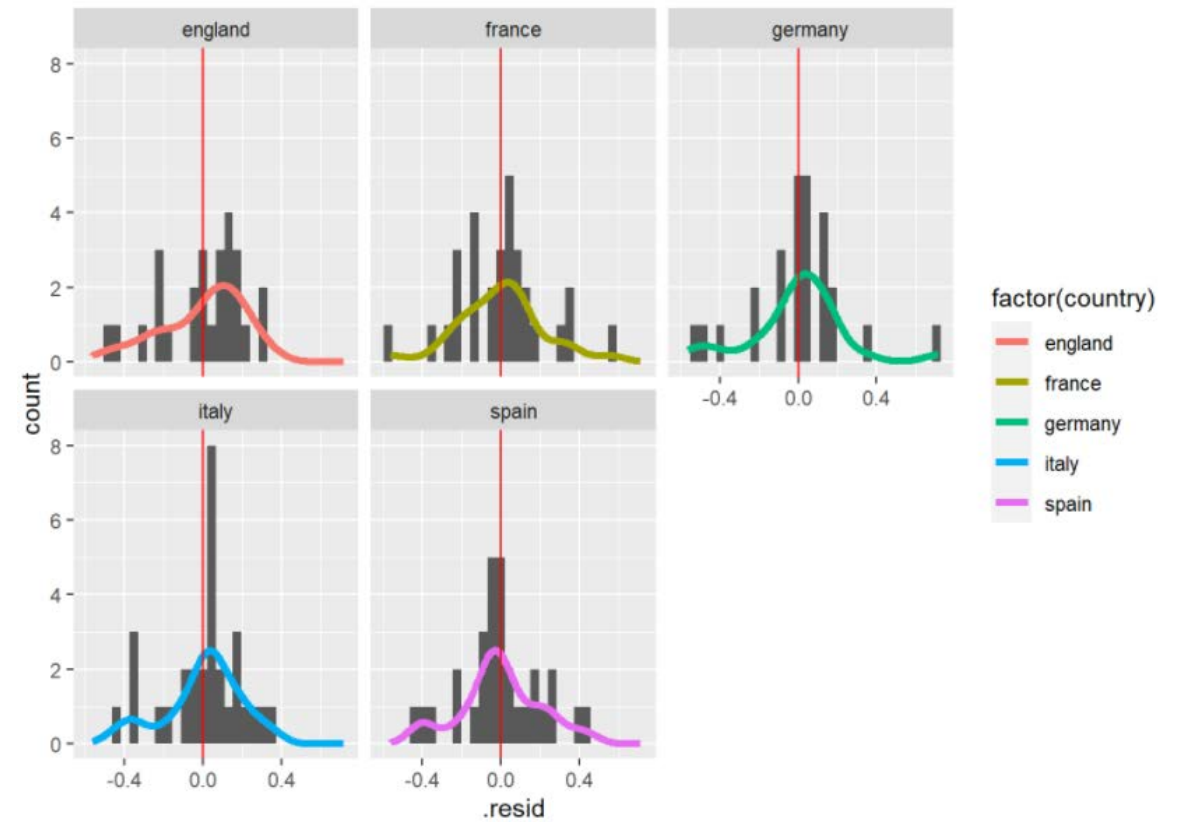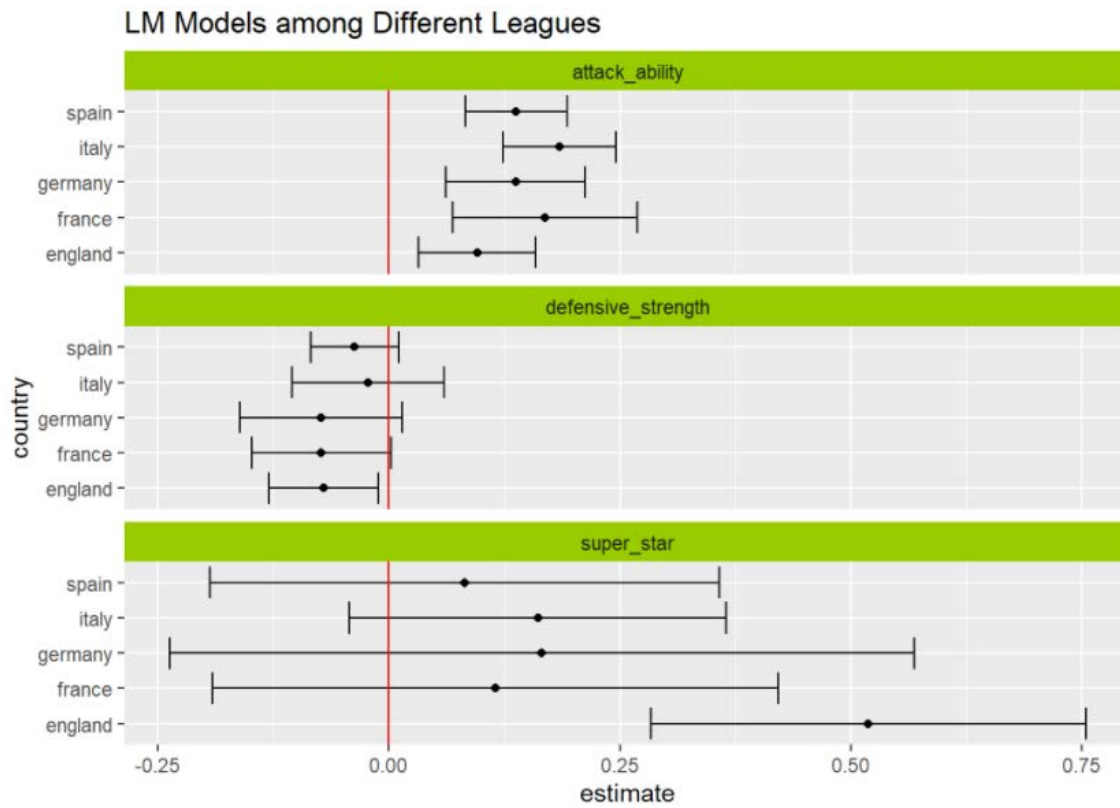
Predict team's performance


```
## 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
```

Fit model to each league

The end

2021/5/5