Soccer metrics: expected goals

Michael Lopez, Skidmore College

Overview

In this lab, we'll look at women's world cup data.

```r
library(RCurl)
library(tidyverse)
url <- getURL("https://raw.githubusercontent.com/statsbylopez/StatsSports/master/Data/sb_shot_data.csv"
wwc_shot <- read.csv(text = url)
names(wwc_shot)
```

Better shot maps

ggplot() has ample ways to enhance shot maps. Consider the following maps

```r
wwc_shot <- wwc_shot %>%
  mutate(is_goal = shot.outcome.name == "Goal")

usa_shot <- wwc_shot %>%
  filter(possession_team.name == "United States Women's")


p1 <- ggplot(usa_shot, aes(location.x, location.y)) +
  geom_point()

p2 <- ggplot(usa_shot, aes(location.x, location.y, colour = is_goal)) +
  geom_point()

p3 <- ggplot(usa_shot, aes(location.x, location.y,
      colour = is_goal, size = shot.statsbomb_xg)) +
  geom_point()
```

1. What features are apparent in p2 that aren't apparent in p1? What features are apparent in p3 that aren't apparent in p2.
2. The following contour plot creates lines where the team has shot in highest densities. The inside line is most dense, corresponding to the center of where the team took shots. What features are apparent in p4 that aren't apparent in p3? What is apparent in p3 that isn't in p4?

```r
p4 <- wwc_shot %>%
  ggplot(aes(location.x, location.y)) +
  stat_density_2d()
p4
```

3. Find another team wwc_shot %>% count(possession_team.name) and plot their shots. How do they compare to the USA Women's team?

Goals versus expectation

4. Let's investigate finishing ability on the USA team. Calculate the total number of goals scored by each player. Who actually scored the most goals?

5. Calculate the number of expected goals scored by each player. Who was expected to score the most goals?

6. The code below (using `group_by(), summarise(), mutate()`), calculate the performance above/below expectation for each member of the USA team who took a shot. Who performed better than expectation? Below? What does the overall distribution say about the USA team?

```
usa_shot %>%
  group_by(player.name) %>%
  summarise(xg = sum(shot.statsbomb_xg),
            g = sum(is_goal),
            n_shots = n()) %>%
  mutate(ou = g-xg) %>%
  arrange(ou)
```

7. Annotate each line of code above to identify what it's doing.

Practice with dplyr

8. For each USA shooter, average the `TimeInPoss` and `DefendersBehindBall` when they took their shot. Filter to make sure you are only looking at players with at least 10 shots. What does this say about how players took shots?

9. Among all players `wwc_shot`, identify the player who finished with the *most* and *least* goals above expectation.

10. Among all goalies (`player.name.GK`), identify the goalie who finished with the *most* and *least* goals allowed above expectation.

11. Among all players, identify the player who took the most headers (`shot.body_part.name == "Head"`).

12. Among all players, estimate the goal rate given different `shot.technique.name`. Which of these tends to lead to have the highest chance of success?

Exploration

A soccer coach wants to know the best places to shoot from. What would you tell the coach? Create a grid across the field using the `cut()` command (for both x and y), and then, within each location, estimate the goal rate. Next, use `geom_tile()` to make a map of goal rates within each cell of the grid you created. For a reminder on `cut()`, see our notes on Hosmer-Lemeshow, or `?cut()`.