Problem set 5: Curry shots

S470/670 Fall 2021

Upload your submission to Canvas by 11:59 pm, Friday 29th October.

The NBA player Stephen Curry attempted 1,598 field goals during the 2015-16 season; of these, 805 were successful. The file `curry2015-16.txt` contains data downloaded from the NBA on all 1,598 of these shots. (Also on Canvas is the file `get-curry.R` contains the script I used to get the data via JSON, but it doesn't work anymore.) Our data set contains 24 variables. The ones we'll pay particular attention to are:

- `EVENT_TYPE`: either "made shot" (successful) or "missed shot" (unsuccessful.)
- `SHOT_MADE_FLAG`: numerical version of `EVENT_TYPE`.
- `SHOT_DISTANCE`: distance from basket in feet.
- `LOC_X`: "horizontal" distance from basket, in tenths of feet.
- `LOC_Y`: "vertical" distance from basket, in tenths of feet.

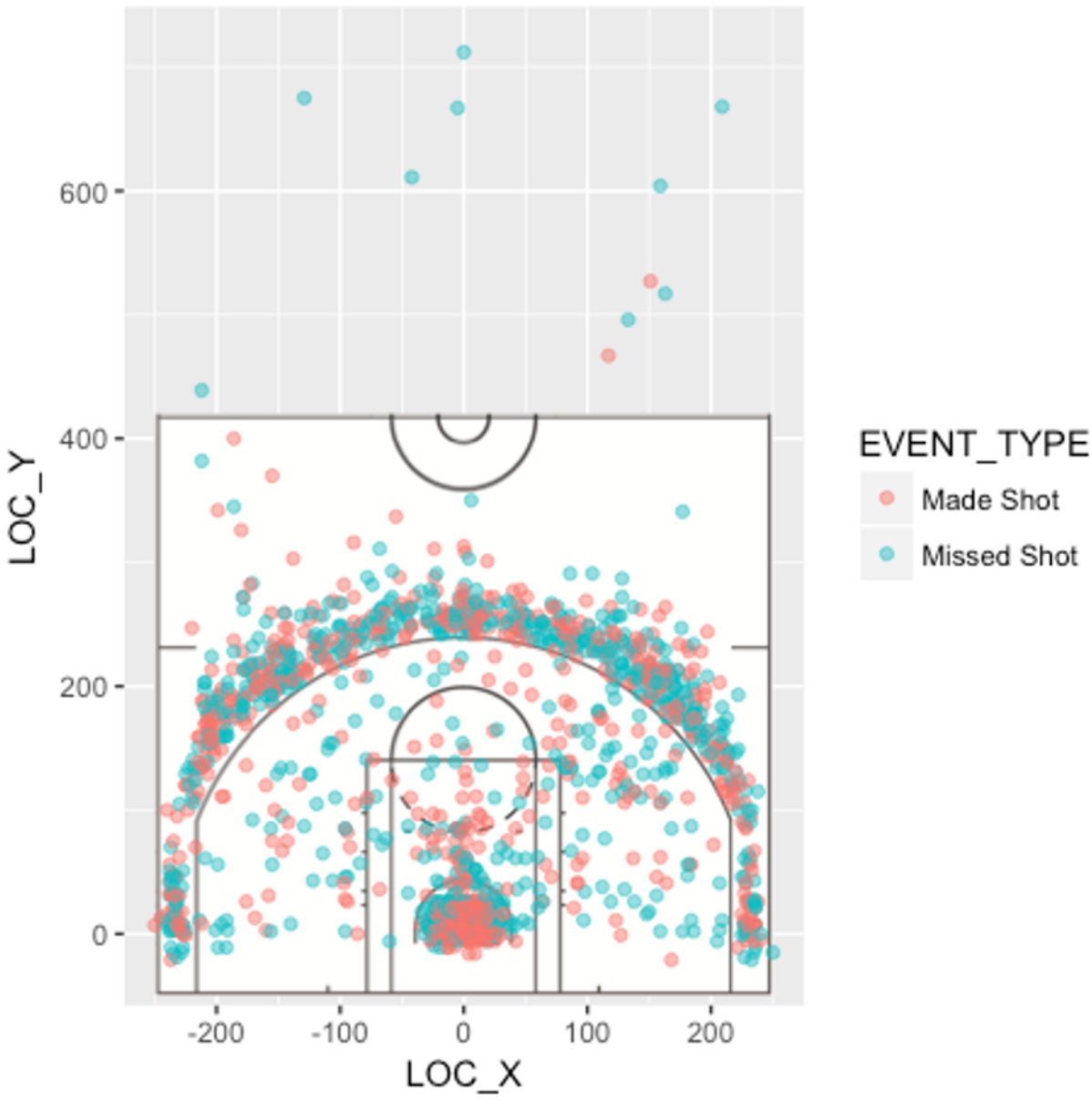
Questions

1. Plot the data to show the location of Curry's shots using color to distinguish between made and missed shots, similarly to the picture below but more colorblind-friendly. (You don't have to include the picture of the court unless you want to show off.) NB: It should use `coord_fixed()` since the units are the same for both axes.

2. Fit a logistic regression to predict whether the shot is made, using the single predictor `SHOT_DISTANCE`. Draw an appropriate ggplot of the fitted curve and write an equation for the fit.

3. Plot the residuals in a way that shows where the logistic regression doesn't fit the data well. Describe in some detail how the model is inaccurate.

4. Fit a better model. You could try a different functional form or a model with more predictors (as long as you use the predictors sensibly.) Your model doesn't have to be perfect, just better. Draw a graph that shows how your model differs from the simple logistic regression, and convince us that your model is better.


600
400
200
0
LOC_Y
-200
-100
0
100
200
LOC_X
EVENT_TYPE
Made Shot
Missed Shot