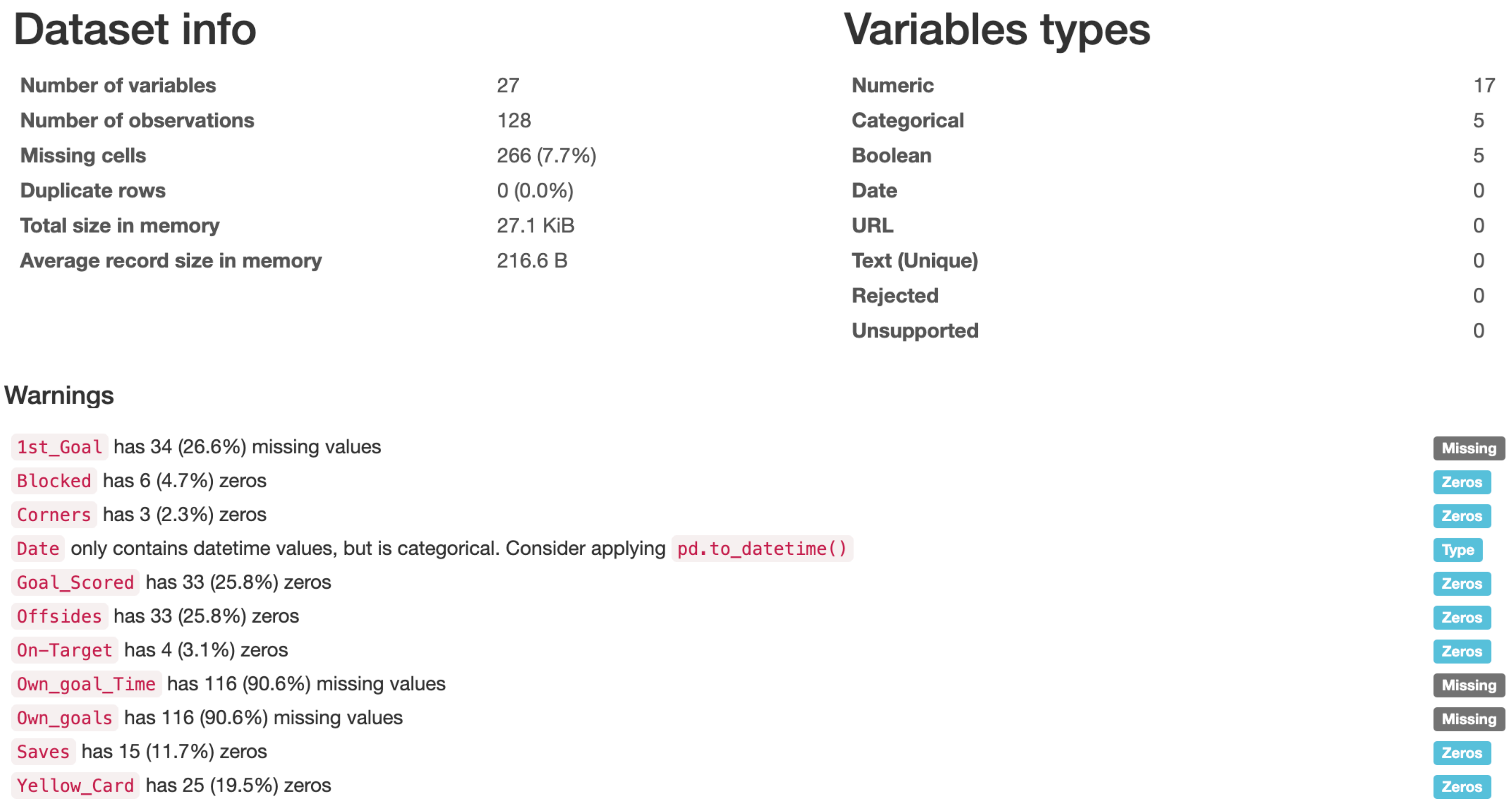
INSAID ML Project: FIFA 2018 World Cup

DataSet Information



Dataset info

Number of variables	27
Number of observations	128
Missing cells	266 (7.7%)
Duplicate rows	0 (0.0%)
Total size in memory	27.1 KiB
Average record size in memory	216.6 B

Variables types

Numeric	17
Categorical	5
Boolean	5
Date	0
URL	0
Text (Unique)	0
Rejected	0
Unsupported	0

Warnings

`1st_Goal` has 34 (26.6%) missing values	Missing
`Blocked` has 6 (4.7%) zeros	Zeros
`Corners` has 3 (2.3%) zeros	Zeros
`Date` only contains datetime values, but is categorical. Consider applying `pd.to_datetime()`	Type
`Goal_Scored` has 33 (25.8%) zeros	Zeros
`Offsides` has 33 (25.8%) zeros	Zeros
`On-Target` has 4 (3.1%) zeros	Zeros
`Own_goal_Time` has 116 (90.6%) missing values	Missing
`Own_goals` has 116 (90.6%) missing values	Missing
`Saves` has 15 (11.7%) zeros	Zeros
`Yellow_Card` has 25 (19.5%) zeros	Zeros

EDA - How to win a World Cup Match

1. Distance Covered - Team Vs Opponent

2. Goal Difference - Team Vs Opponent

3. Fair Play - Foul Rate

4. Precision - Pass Accuracy & Shot Accuracy

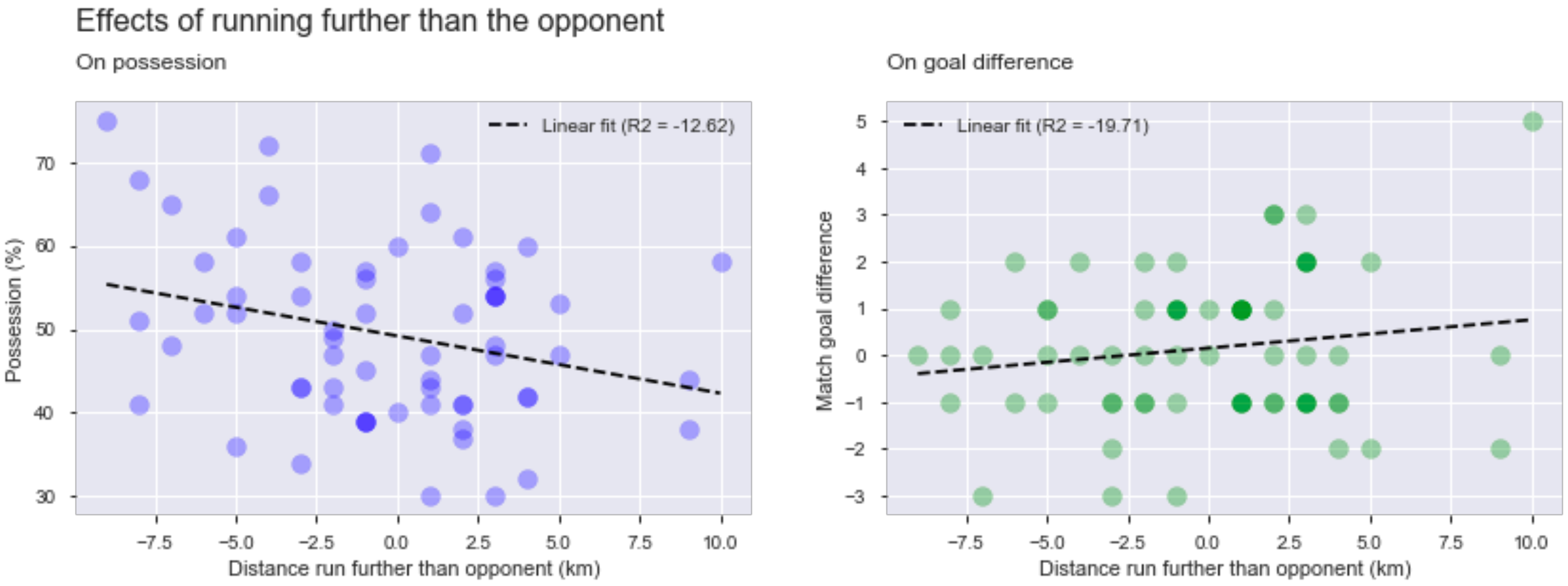
5. Be South American / European? (It illustrates a trend common across World Cups: that the South American and European teams usually do well.)

1. Distance Covered - Team Vs Opponent

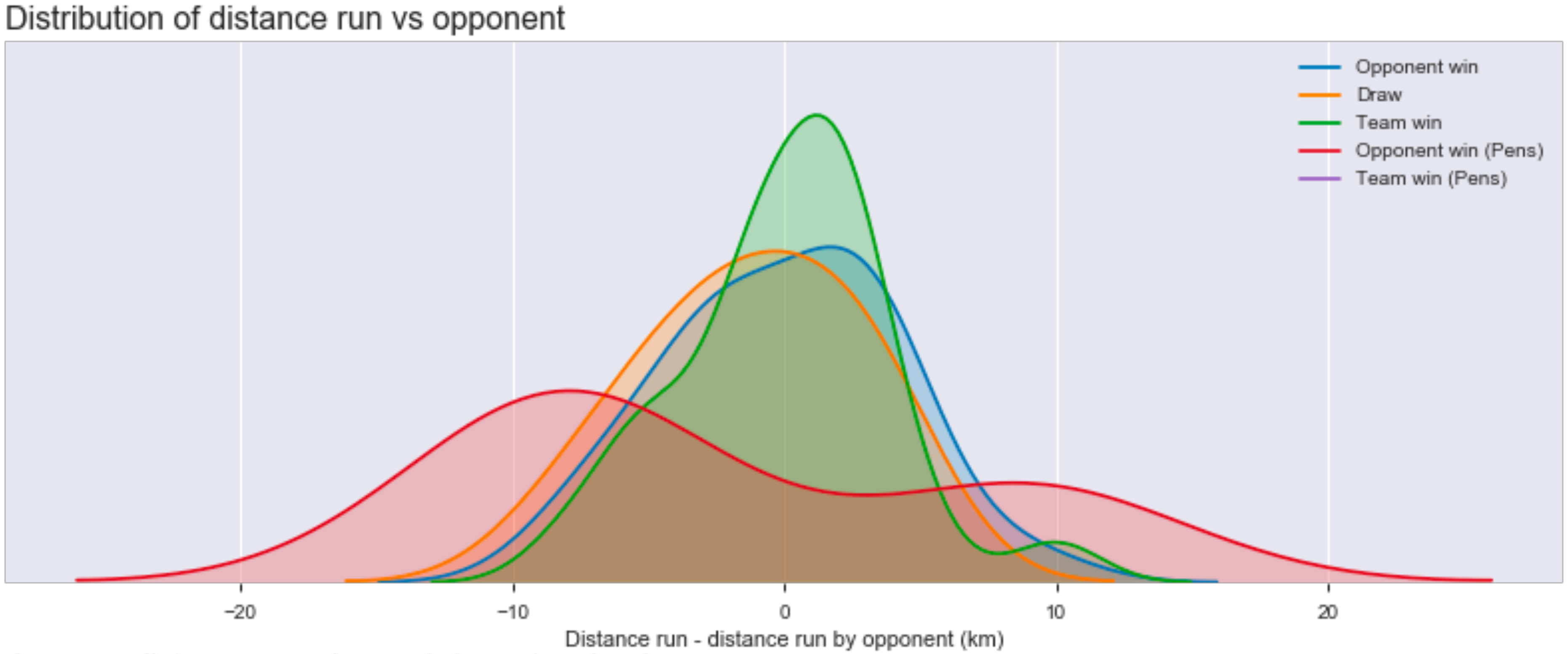
- Running more suggests that possession will be lower - which makes sense since teams generally have to do more running out of possession to get in defensive position, cover possible passes

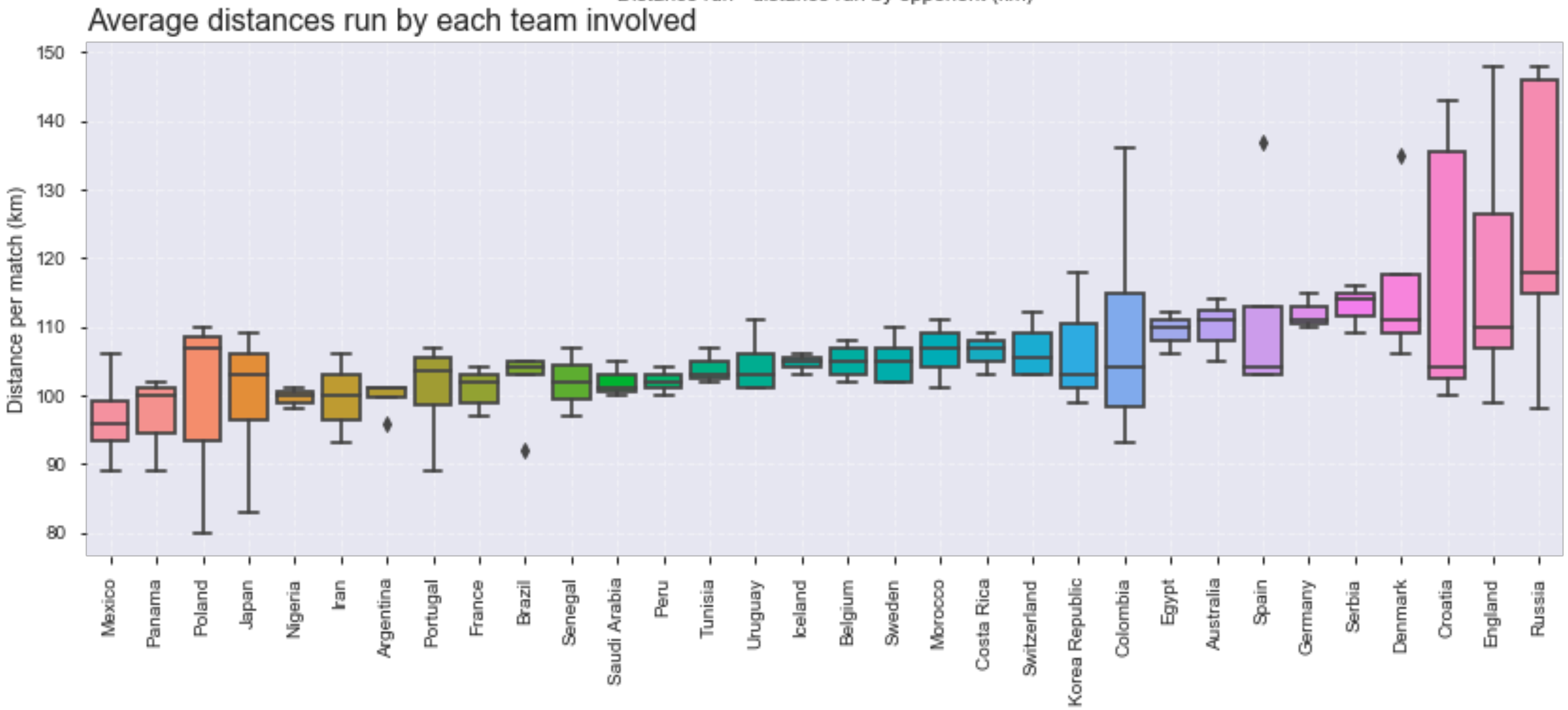
2. Goal Difference - Team Vs Opponent

- Goal difference in the match is less clear. Visually there seems to be a positive correlation but the r2 score isn't great.



3. Distribution of Distance Run - The eventual winners, France, are towards the lower end of distance run - whilst two of the semi finalists, England and Croatia, were some of the hardest runners.

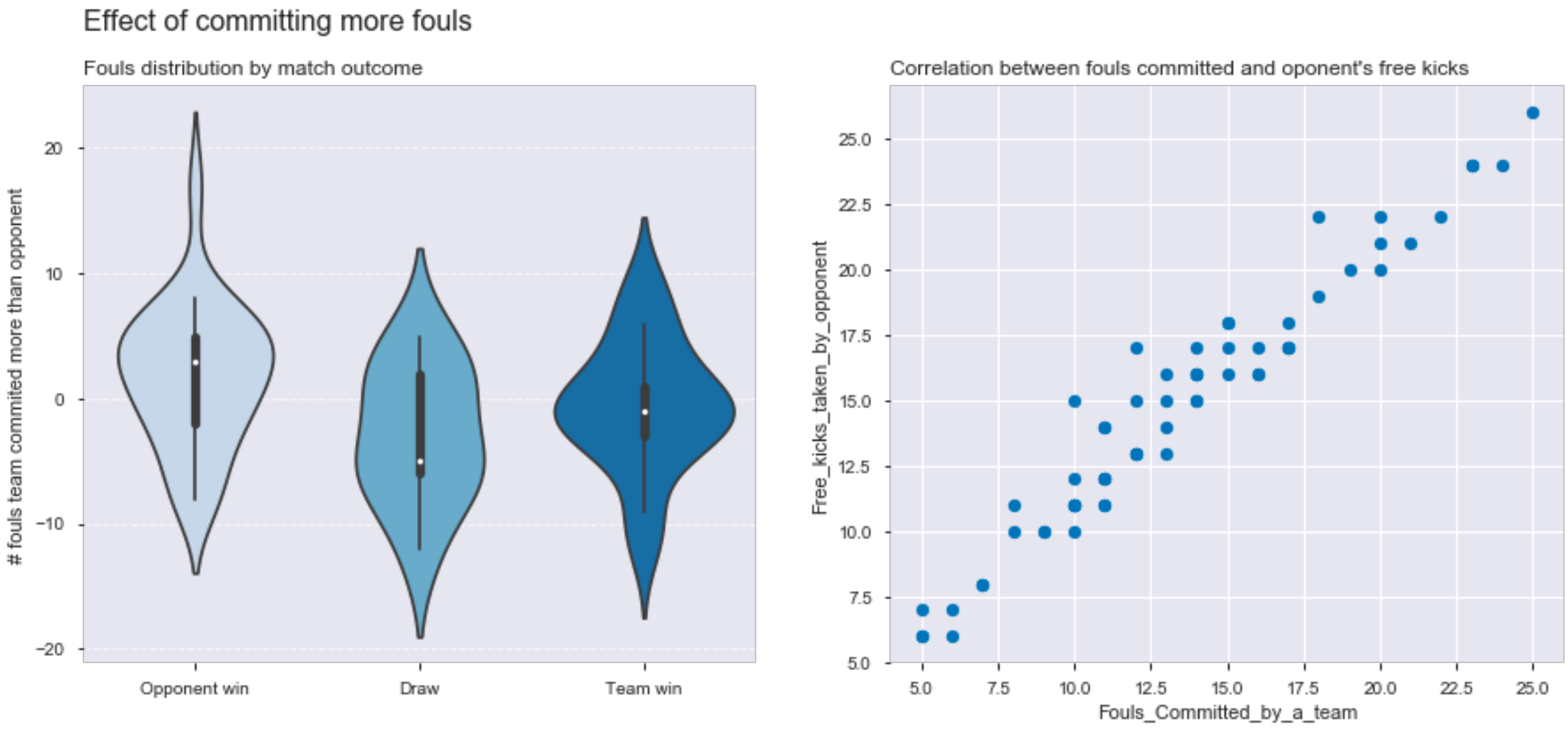

Distribution of distance run vs opponent


Average distances run by each team involved

4. Fair Play - (Fewer fouls committed = Better chance of winning.)

Again we can't be sure if this is a correlation or a causation - losing teams are more likely to foul out of desperation, and in doing so concede set pieces.

What we do know from the second plot, which in retrospect is no surprise at all, is that it definitely leads to the opposition having more free kicks. Tricky to pin down

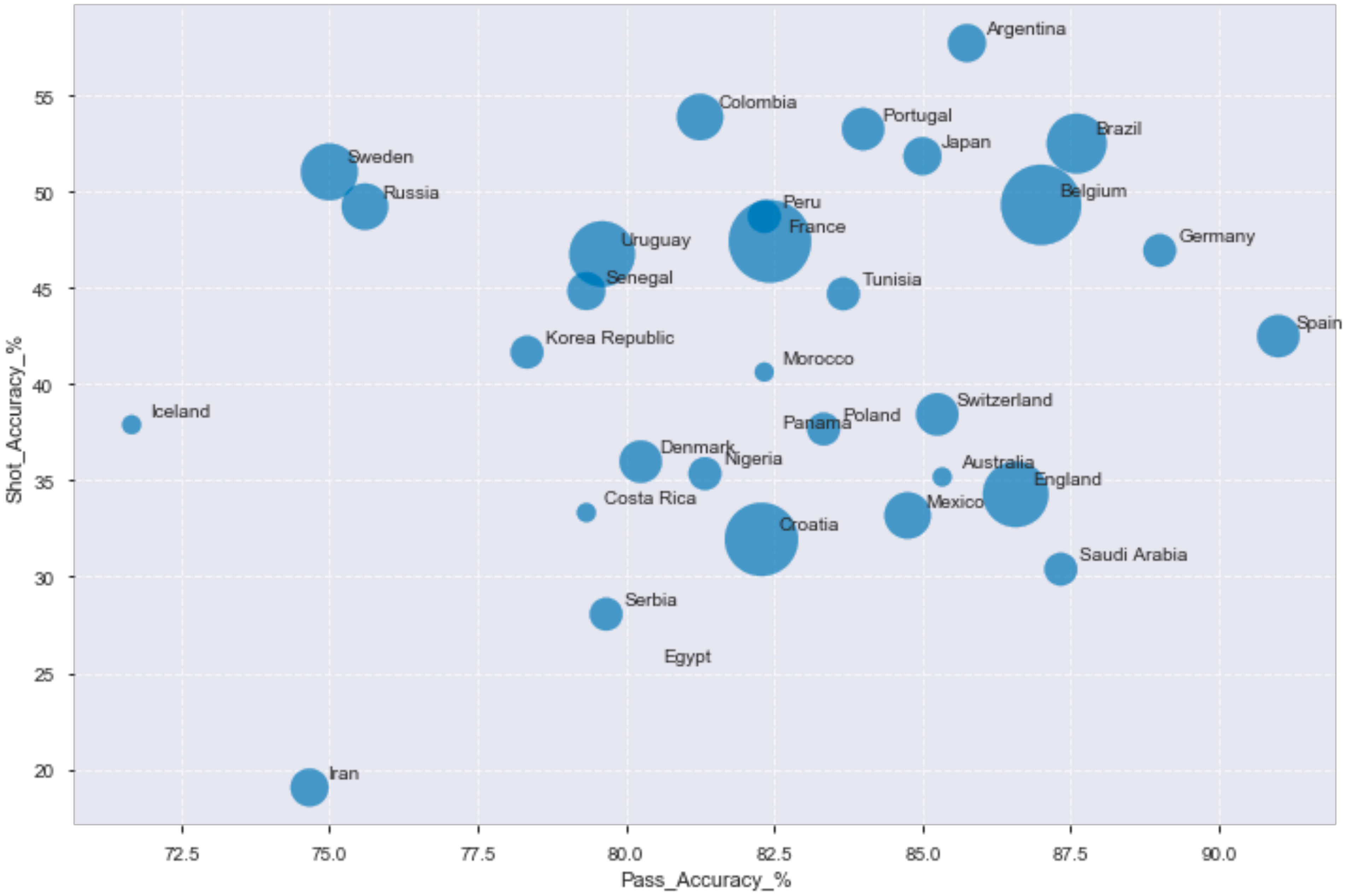

Effect of committing more fouls

Fouls distribution by match outcome

Correlation between fouls committed and oponent's free kicks

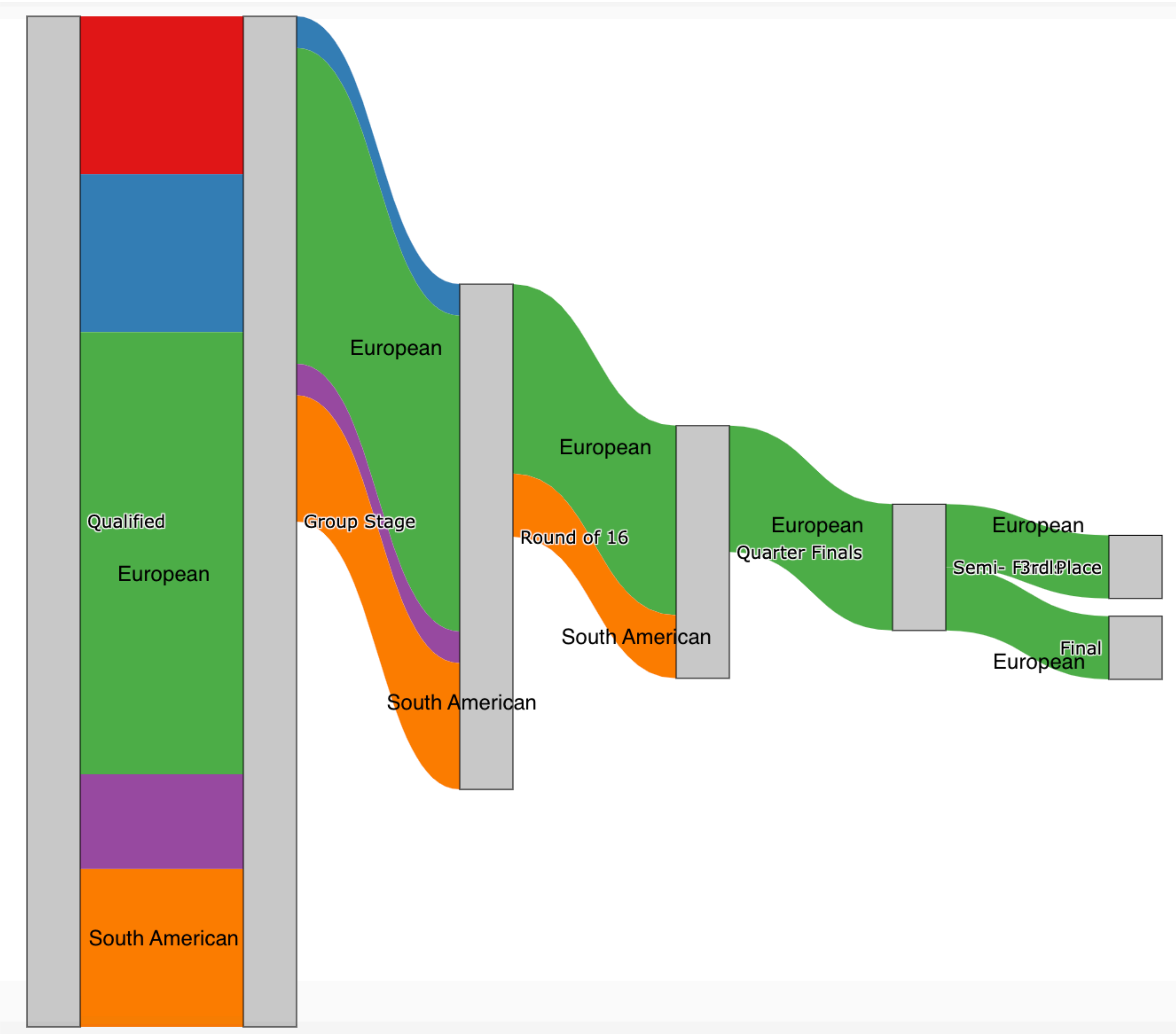
5. Precision - (Shot Accuracy and Pass Accuracy)

- If does seem that shot and pass accuracy are correlated, but it is not as clear whether they both increase performance (marker size).
- It appears that pass accuracy has a larger effect than shot accuracy if anything.
- Spain killing it on the pass accuracy as usual, whilst Iran taking pot shots apparently



6. Be South American / European?

It illustrates a trend common across World Cups: that the South American and European teams usually do well.

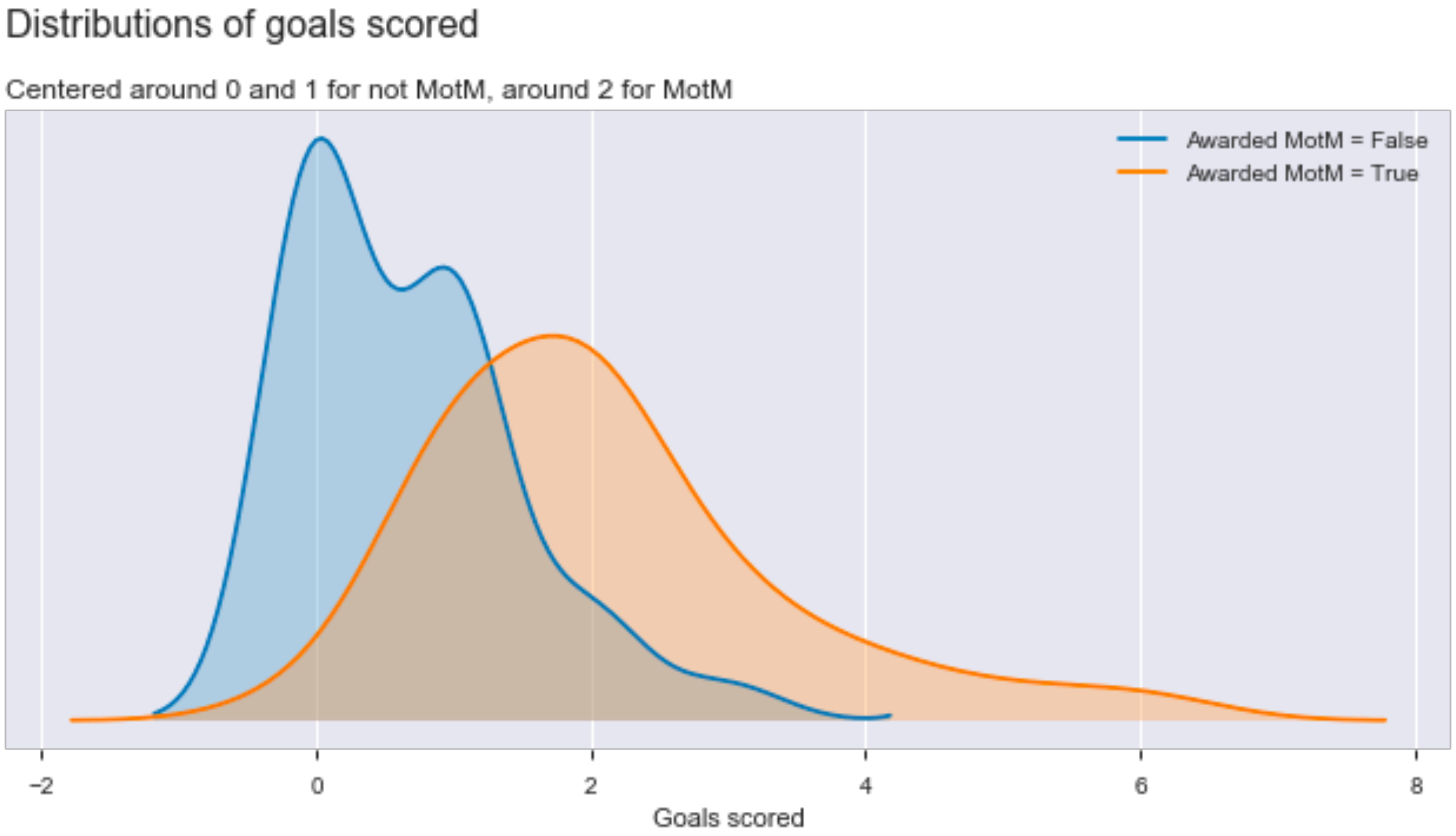


EDA - Explore relationships with MotM

1. Team goals scored vs. MotM
2. Team possession vs. MotM
3. Match outcome vs. MotM

1. Team goals scored vs. MotM

The more goals a team scores the more likely the MotM is to be on their team.

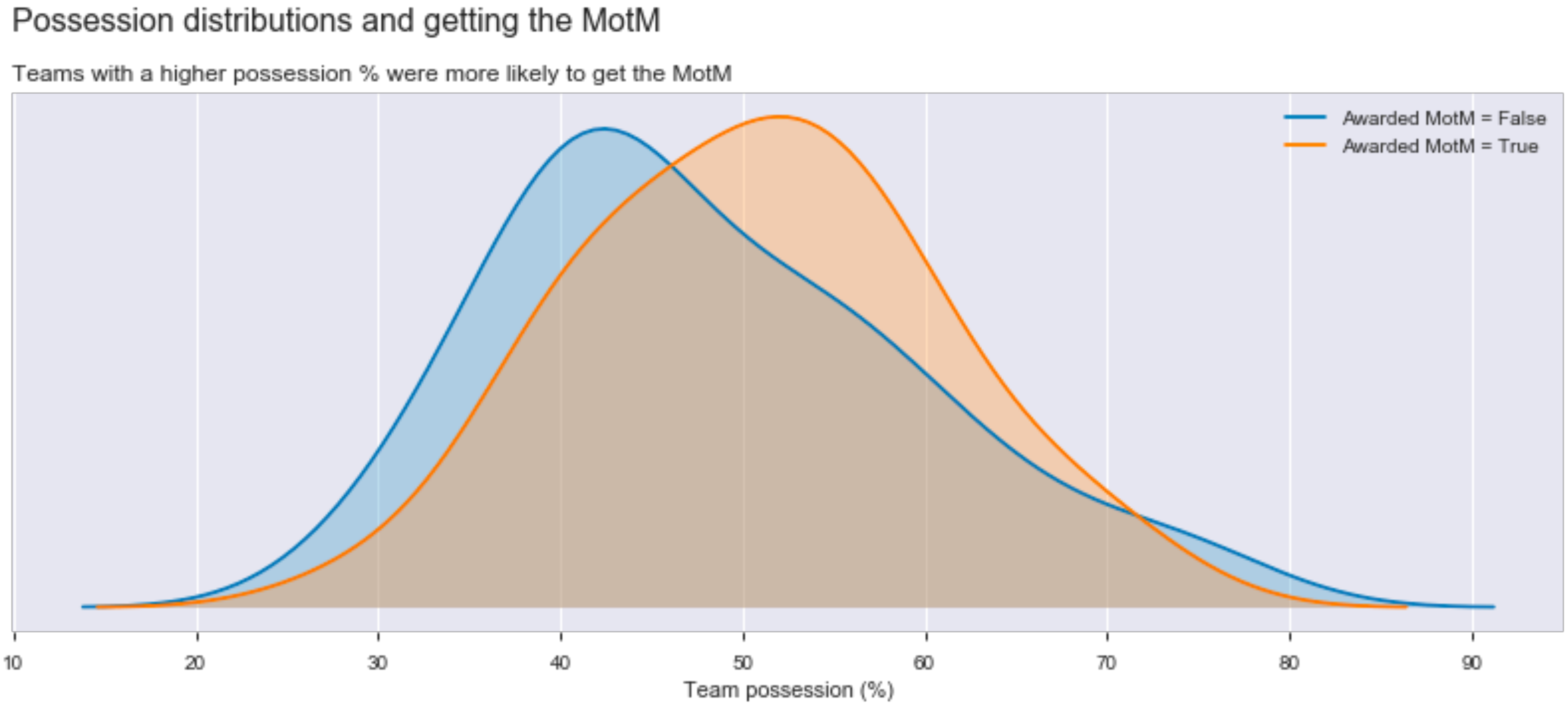


Distributions of goals scored

Centered around 0 and 1 for not MotM, around 2 for MotM

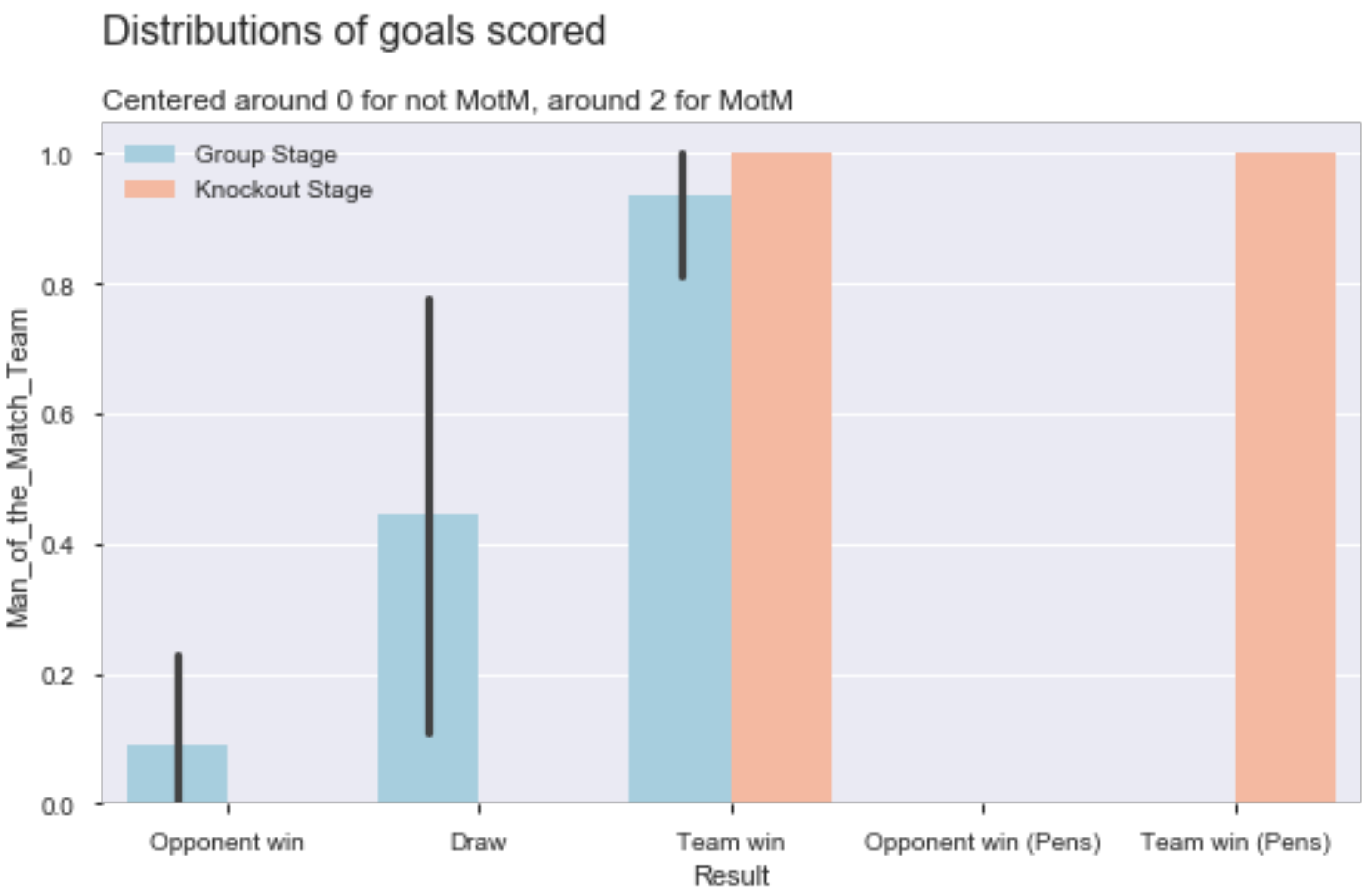
2. Team possession vs. MotM

Higher team possessions lead to a reasonably higher chance of getting MotM, but not hugely so.



3. Match outcome vs. MotM

- In the group stages, the MotM is virtually always on the winning team with a few rare occurences on the drawing and losing sides.
- In the knockout stages, the MotM is exclusively on the winning side



ML - Predict MotM

1. Approaches (using Original Dataset and optimised EDA dataset)

2. Logistic Regression, KNN and Random Forest Algorithm

3. Output Classification Report

4. Feature Importance for RF

ML - Appraoch

- Missing Data Treatment

- Remove redundant features

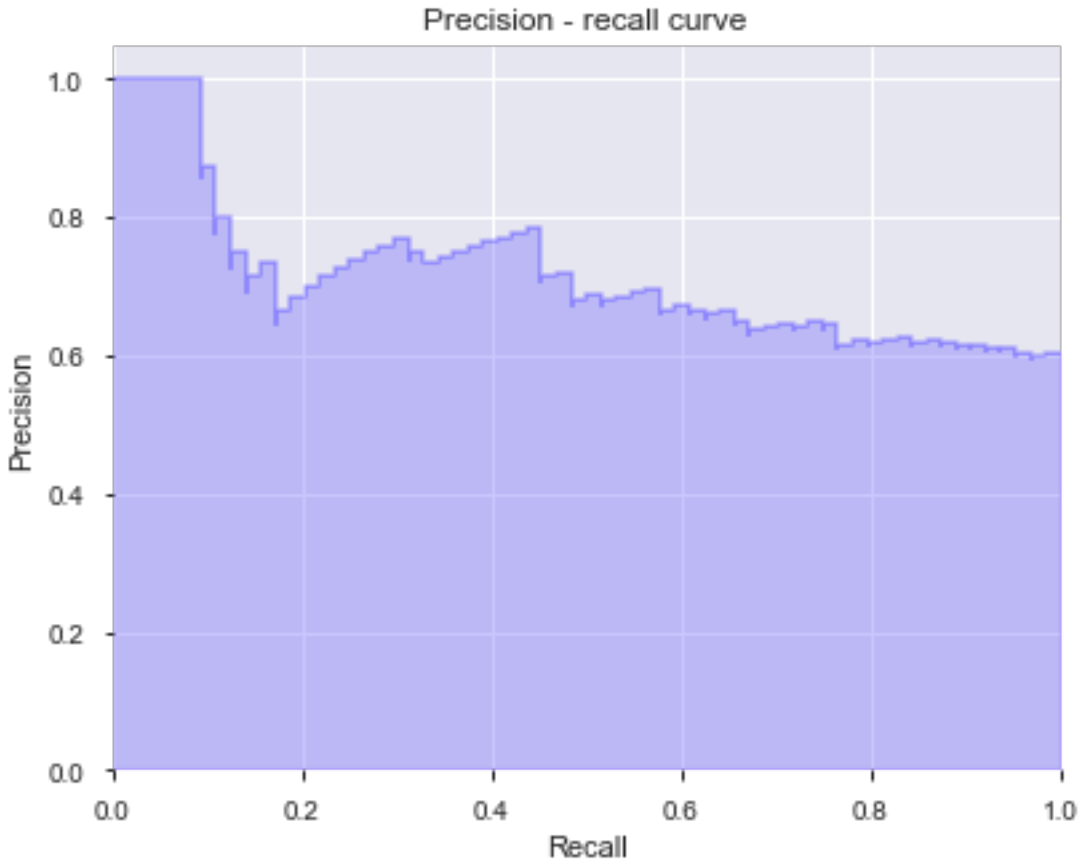
- Encoding & dummify categorical variables

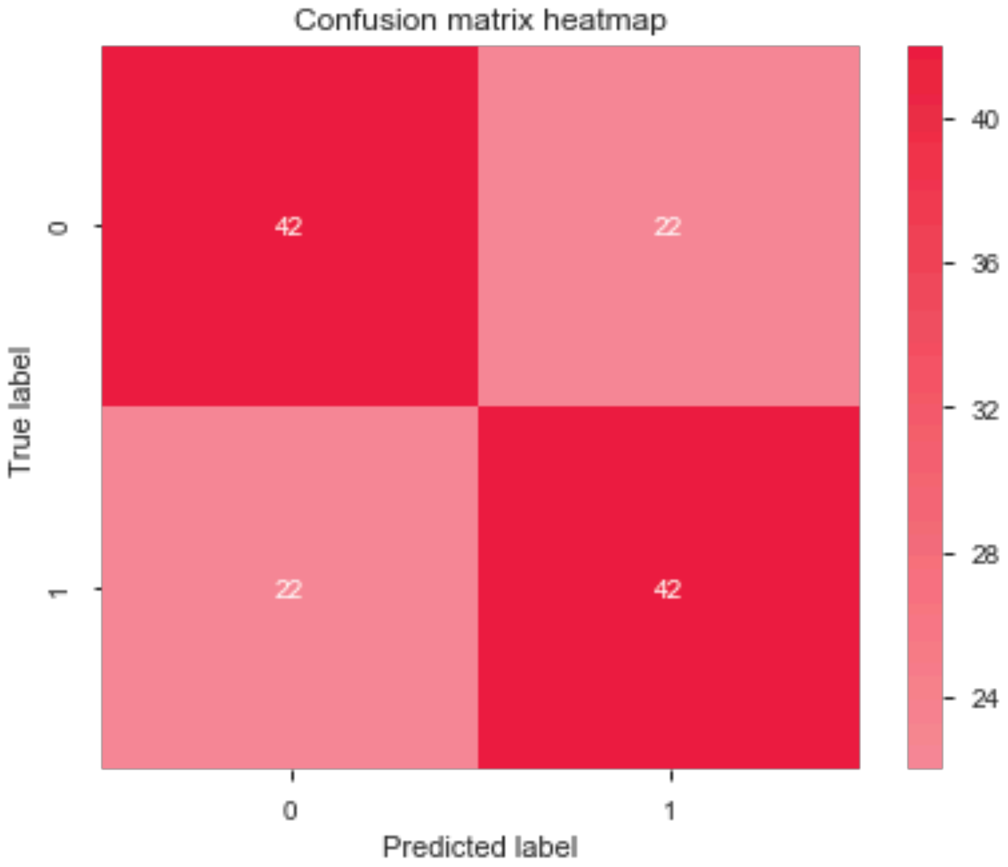
- Standardise the features

- Split in 2/3rd and 1/3rd as train and test

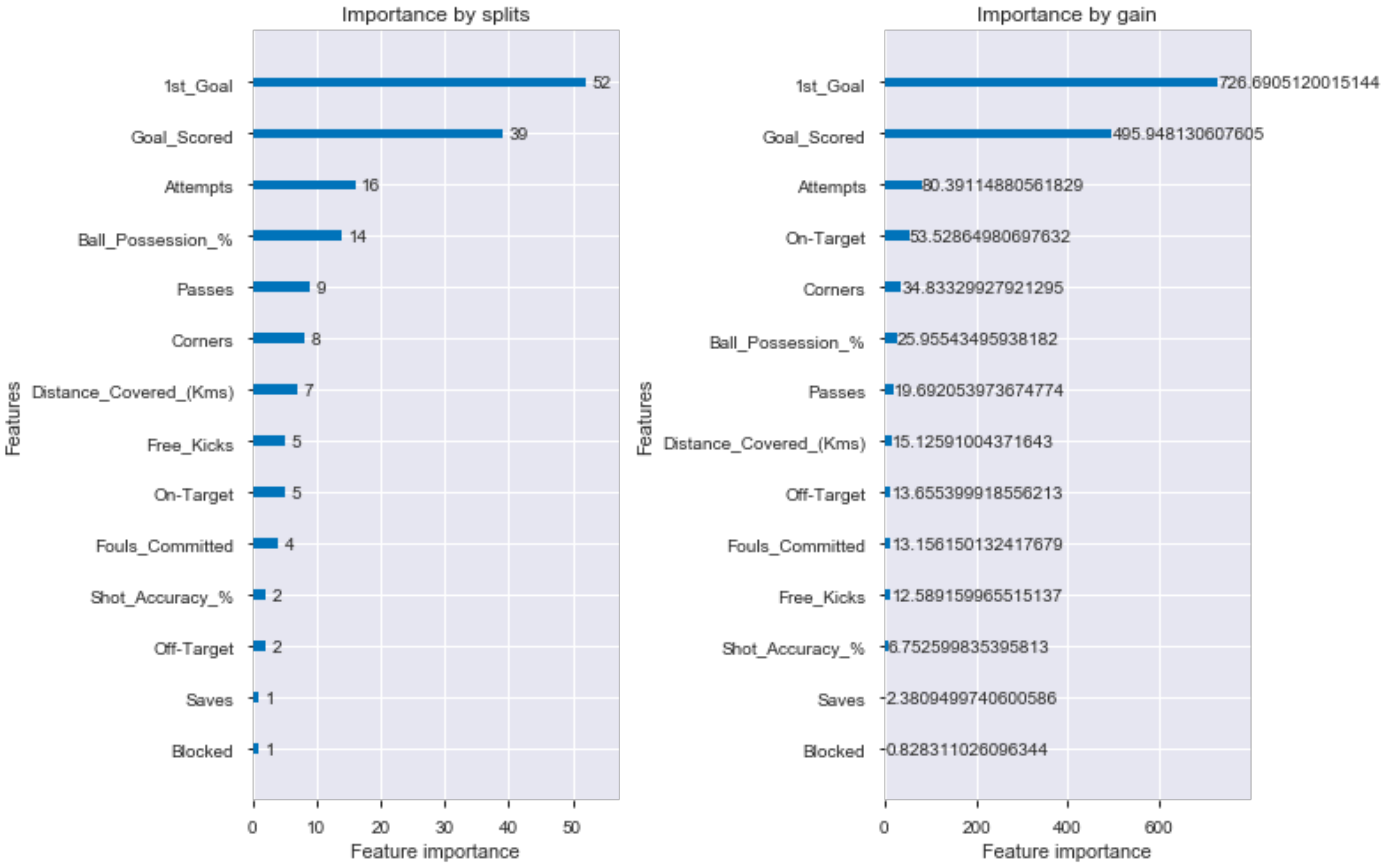
- Classify over LogR, KNN & RF

3. Results


Precision - recall curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Confusion matrix heatmap
True label
Predicted label
0
1
42
22
22
42
0
1
40
36
32
28
24

3. Feature Importance


Importance by splits
Features
1st_Goal
Goal_Scored
Attempts
Ball_Possession_%
Passes
Corners
Distance_Covered_(Kms)
Free_Kicks
On-Target
Fouls_Committed
Shot_Accuracy_%
Off-Target
Saves
Blocked
Feature importance
Importance by gain
Features
1st_Goal
Goal_Scored
Attempts
On-Target
Corners
Ball_Possession_%
Passes
Distance_Covered_(Kms)
Off-Target
Fouls_Committed
Free_Kicks
Shot_Accuracy_%
Saves
Blocked
Feature importance