Team Hyperborea : Sports | Individual Milestone 1

Annapoorani Sundararaj Shanthi

When it comes to sports, particularly soccer, home and away advantage can be a significant factor in many of the results and statistics of the match. For instance, a home side could have a home advantage, and the referee could be in favour of them, and that might lead to fewer fouls and penalties being brought up against them, and more for the away team. Many players, including us at some point, have said that they choose to play at home because of the environment and not have to care about commuting and being tired from going to the field.

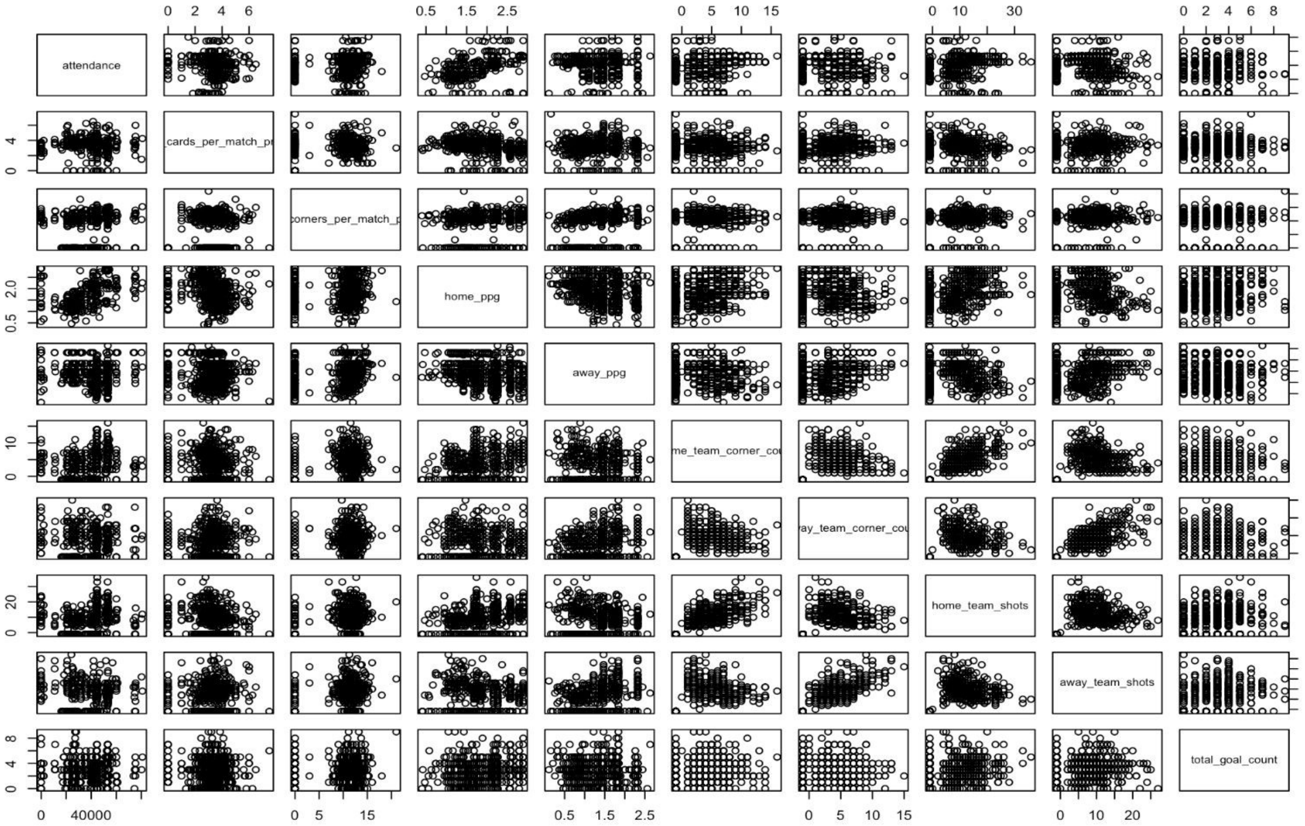
"Home field advantage" is a sporting term that has links to the fan base's feelings and anecdotal encouragement. We want to be able to report whether the evidence confirms this effect, but also which statistics forecast this difference, to point a casual fan's eye to a data-based "why." It may also help advise people when to tune in to a team that they could only casually follow. If this disparity is pronounced enough, a fan may be encouraged to stop attending games in which Liverpool is playing-because the outcomes will be disappointing and thus not fun for that fan.

The project will explore whether or not there is a benefit to play at home or away with the Liverpool European Soccer Team. Specifically, we expect to answer an experimental query with a null hypothesis that models return home statistics when Liverpool is playing home that are equivalent to away statistics when Liverpool is away, with an alternative hypothesis that our models return home statistics that are better than away statistics for a range of response. Using the EPL Matches results, we hope to address a few questions about the relationship between home field advantage or away field disadvantage and goals scored per game. In this report, I focus specifically on Possession at home or away and the xG (Expected goals).

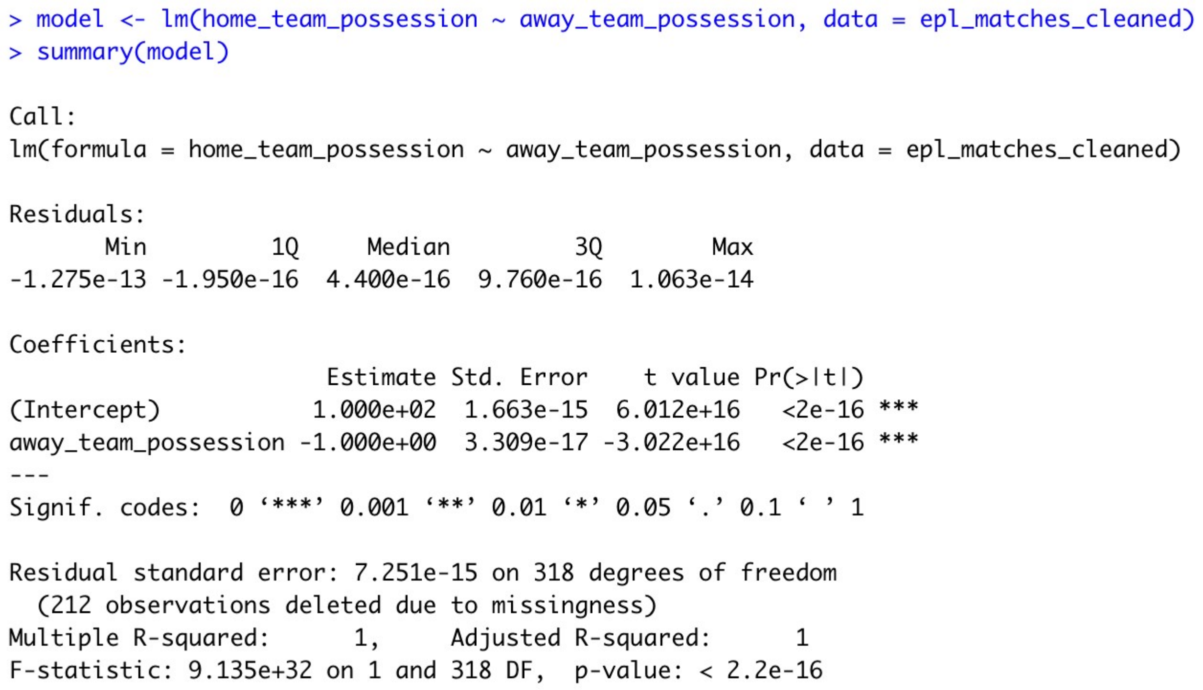
Possession means that a team or player is in control of the ball. A player (and his team) are in possession while he has the ball, or is passing to or receiving from another member of his team. If he loses the ball, he loses possession. One of the best things about having good possession numbers is how much it can frustrate the opposing team.

Not only do teams get frustrated with the fact that they can't get the ball, they will often make multiple mistakes in an attempt to regain possession. The **response variables will be home_team_red_cards, away_team_red_cards and home_team_yellow_cards, away_team_yellow_cards.** I have built models that focus on possession correlation of home and away.

xG is basically expected goals measures the quality of shots on several variables such as assist type, shot angle and distance from goal, whether it was a headed shot and whether it was defined as a big chance. Here there are two teams, home team and the away team, the xG which is the expected goals scored by each team, the team with higher expected goals makes the strong correlation. To explore this part of the dataset, I have created models displaying that correlation. The focal point of the model is that it is the best single metric for understanding performance of a soccer team and predicting it's future. **The response variables will be home_team_shots_on_target and away_team_shots_on_target and also total_goals_at_half_time.**

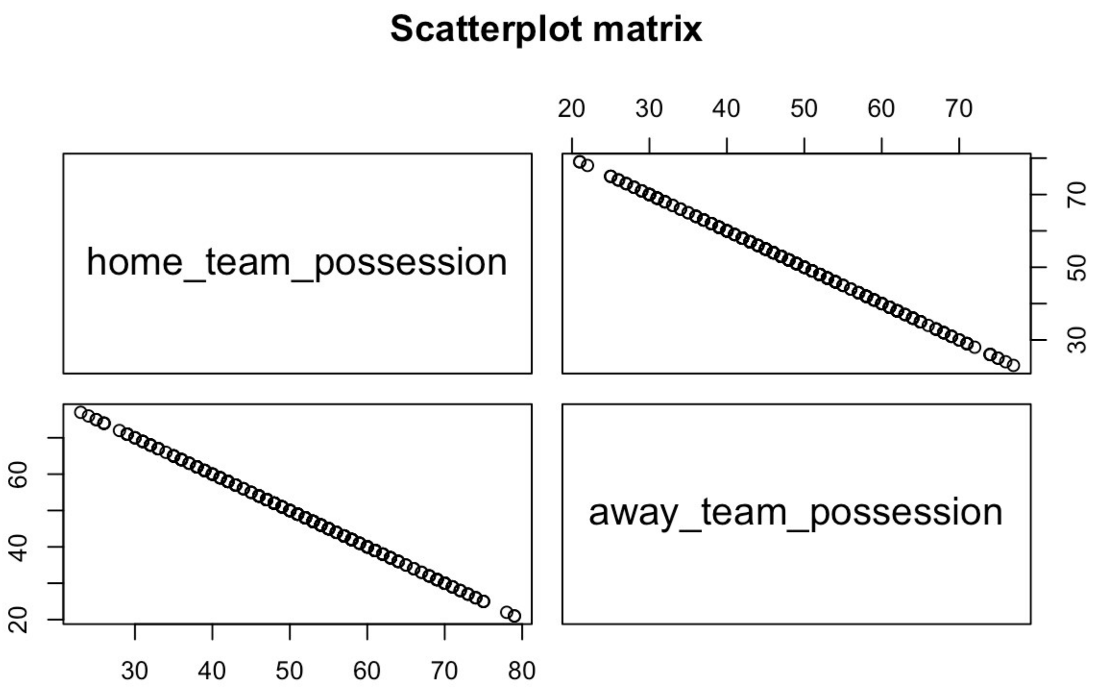


From the above plot, we could find that there's not high correlation between any two independent variables which implies that there’s no multicollinearity exists between them. First, we build a model all the first order terms and it passes the F-test wherein at least one of the Beta value is not equal to null.

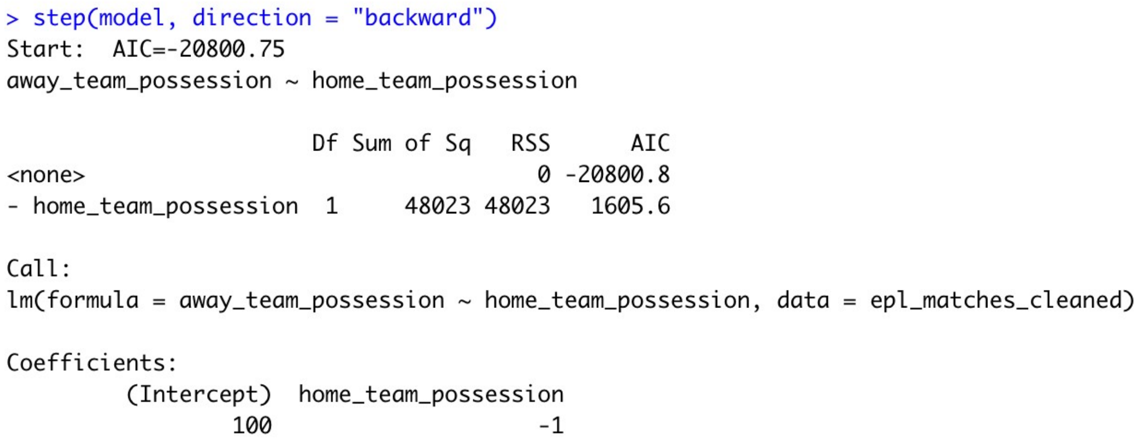


```
> model <- lm(home_team_possession ~ away_team_possession, data = epl_matches_cleaned)
> summary(model)

Call:
lm(formula = home_team_possession ~ away_team_possession, data = epl_matches_cleaned)

Residuals:
       Min         1Q     Median         3Q        Max 
-1.275e-13 -1.950e-16  4.400e-16  9.760e-16  1.063e-14 

Coefficients:
                      Estimate Std. Error    t value Pr(>|t|)    
(Intercept)          1.000e+02  1.663e-15  6.012e+16   <2e-16 ***
away_team_possession -1.000e+00  3.309e-17 -3.022e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.251e-15 on 318 degrees of freedom
  (212 observations deleted due to missingness)
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 9.135e+32 on 1 and 318 DF,  p-value: < 2.2e-16
```

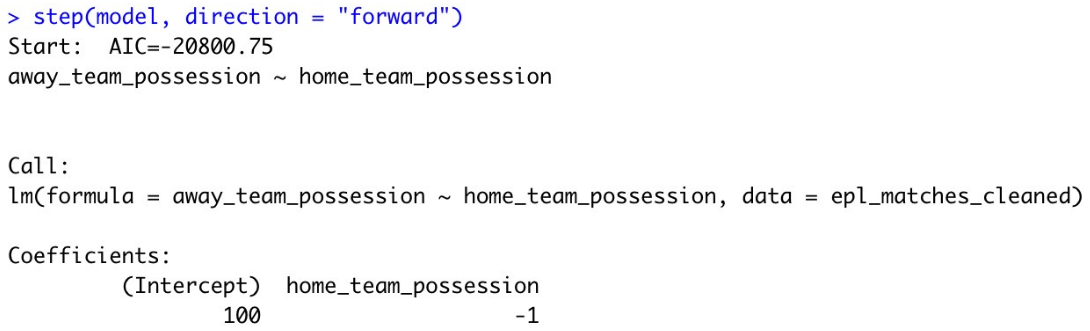
The scatterplot matrix for possessions is,



I tried Forward and backward selection for home team possessions and away team possessions, the results are



```
> step(model, direction = "backward")
Start:  AIC=-20800.75
away_team_possession ~ home_team_possession

                       Df Sum of Sq   RSS      AIC
<none>                                  0 -20800.8
- home_team_possession  1     48023 48023   1605.6

Call:
lm(formula = away_team_possession ~ home_team_possession, data = epl_matches_cleaned)

Coefficients:
         (Intercept)  home_team_possession  
                 100                    -1  
```



```
> step(model, direction = "forward")
Start:  AIC=-20800.75
away_team_possession ~ home_team_possession


Call:
lm(formula = away_team_possession ~ home_team_possession, data = epl_matches_cleaned)

Coefficients:
         (Intercept)  home_team_possession  
                 100                    -1  
```

After seeing the results from the above experiments, I guess team which plays in their home ground will have higher probability to win the game compared with the away team. That's why it is called "Home field advantage". Here that term is experimentally proven with the experiments given.

The same experiment has been done to xG as well to predict the expected goals of the two team, home team and the away team. As proven earlier, the team which played in their home i.e. Home team, will have the higher chance of scoring more goals than the away team. The focal point of the model is that it is the best single metric for understanding performance of a soccer team and predicting it's future.

For the next milestone, I'll add in second order terms for the possessions and xG to see if there are any changes in correlation and if the away team has even a slightest chance of winning the game.