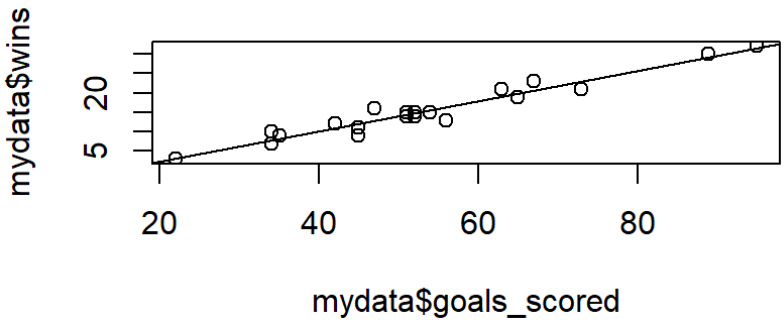
Soccer Analysis

The purpose of this analysis was to find out a unique statistic or variable of a soccer game, which indirectly affects the chances of winning a game. My thought process was that at first, I will find a primary variable (heaviest correlation with winning a game). Thereafter, I could look for other variables and compare their correlation with the primary variable. My goal was to look for a variable that is overlooked and underrated, but that which indirectly increases the chances of winning a game. Below is the process of how I carried out this analysis. The code will be provided as a R file.

Initially, I looked for the highest correlation between the variable *wins* and other variables. This variable turned out to be *goals scored*, with a correlation of 0.97. This is obvious and it is expected that the chances of winning a game increase with goals scored. By fitting a regression line on top of the scatter plot, and looking at the R squared value, it can be concluded that the chances of winning a game increases with the number of goals scored.




```
Residual standard error: 1.755 on 18 degrees of freedom
Multiple R-squared:  0.9445,    Adjusted R-squared:  0.9414

F-statistic: 306.5 on 1 and 18 DF,  p-value: 9.461e-13
```

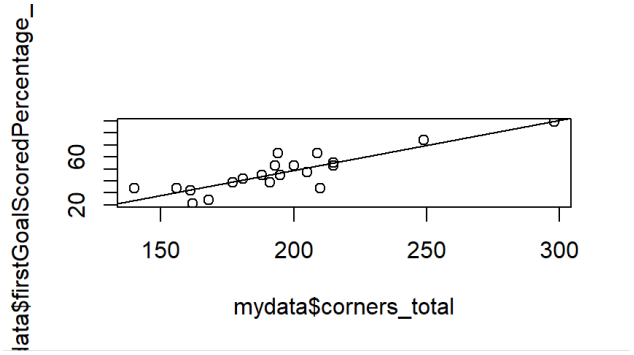
Now that I had established that the number of goals scored is the primary variable, I could compare its correlation with other variables. The reason for this is that, although it is helpful to know that the number of goals scored increases the chances of winning a game, there needs to be a gameplan or some kind of knowledge to be able to achieve it. A team cannot simply decide to score more goals as a game plan. Many teams rely on passing the ball around a lot (average possession) and some teams rely on counter attacks. My aim was to utilize a tactic which takes advantage of an underration correlation between this desired variable and goals scored. This variable could be thought of as an X factor.

I analysed the correlations and found that all the strong correlations between the number of goals scored and the other variables were with expected ones, such as *shots* taken (correlation of 0.84), and *average possession* (correlation of 0.84). However, there was one exception and and also an unexpected one. The correlation between the *number of goals scored* and the *first goal scored perecentage* was 0.93. This was a surprise because this means that as long as a team scored the most number of first goals, there was a good chance that the team would go on to score the highest number of goals in a season. I wanted to quickly verify and see if the the top two teams that highest number of goals scored held the place for first goal scored percentage as well. Sure enough, they did and the teams were Manchester City and Liverpool. These two teams also ended the season as winners and runners up respectively.

This does not mean that scoring the first goal is more important than taking more shots and keeping possession of the ball which had a high correlation as well. These are just correlations which measure the strength of the relationship between two variables, and explains that the variables follow a similar trend in increasing as the other increases (in this analysis).

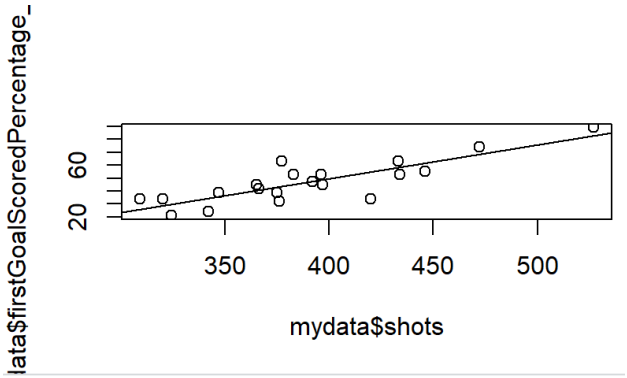
Although the variable *first goal scored percentage* is not directly applicable as a game plan, similar to scoring goals, the path was narrowed to look for another variable, hidden and underrated, which may have a strong correlation to the *first goal scored percentage*. The plan is to look for a variable that can be directed to the players as a game plan to achieve the first goal. Again, I did some correlation analysis and found an unexpected result. The highest correlation existed between *first goals scored percentage* and *total corners* taken (correlation of 0.87), as compared to *average possession* and *shots* taken. The R squared value was best for *toal corners* as well.

One thing to observe is that although the correlations are high, we can see through the R-squared value, that the variations in the data are not explained that well by any of these variables. This is due to the countless variables that go into play due to the chaotic nature of a soccer game which may or may not even be present in the data.



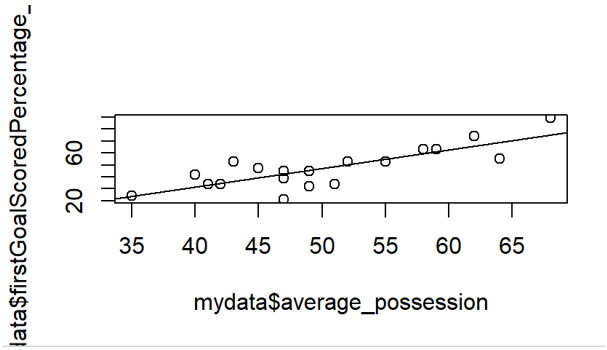

```
Residual standard error: 8.473 on 18 degrees of freedom
Multiple R-squared:  0.755,     Adjusted R-squared:  0.7414

F-statistic: 55.47 on 1 and 18 DF,  p-value: 6.681e-07
```




```
Residual standard error: 9.223 on 18 degrees of freedom
Multiple R-squared:  0.7097,    Adjusted R-squared:  0.6936

F-statistic: 44.01 on 1 and 18 DF,  p-value: 3.16e-06
```




```
Residual standard error: 10.05 on 18 degrees of freedom
Multiple R-squared:  0.6552,    Adjusted R-squared:  0.636
F-statistic:  34.2 on 1 and 18 DF,  p-value: 1.54e-05
```

In a game as chaotic as soccer, there are many factors that go into play in deciding an outcome and it is simply too hard. What I did was take the surface level statistics like *goals scored* which were strongly correlated to many expected variables (shots taken and average possession) and I was able to find out one unexpected variable that is overlooked, which was the first goal scored percentage. From here on, I could further branch deeper to find out which variable had a strong correlation to this underrated and overlooked variable (first goal scored percentage). I found out that it was the total number of corners taken. These two variables had a correaltion of 0.87.

The best teams today and in history, have been those who have been able to keep high possession of the ball. This analysis showed that the correlation between goals scored and the average possession was decent enough (0.84). This strategy is well implemented in a real game, and therefore, it seems like practically applying this strategy based on a correlation of this strength is valid.
Similarly, as implementation of a novel soccer strategy, concentrating on getting corners is a legitimate and very effective way to score the first goal based on its correlation, which I was able to find out. Having a correlation with *goals scored* even higher than *average possession*, it is justifiable to base a teams strategy on winning corners. This can directly lead to scoring the first goal, and thus win more games in a season.