Premier League Analysis

Hovsep Avagyan

The problem/data description

In this project I am going to work with 2 similar datasets. The main objective of gathering these two datasets is the same – to come up with the key factors which identify the final position of the team in British Premier League. The first dataset contains 42 observations in total, and is the data of 14 teams for three seasons (2016/2017, 2017/2018, 2018/2019). The dependent variable is the final position of the team and the independent variables are *Position*, *M_median*, *D_median*, *G_median*, *F_median*, *Shots*, *ShotsOnTarget*, *Falls*, and *Corners*. *Position* is the final position of the team in the Premier League Table, *M_median* is the median height of the midfielders in that team, *D_median* is the median height of defenders, *G_median* is the median of the goalkeepers, *F_median* is the median height of forwards, Shots is the number of shots taken by a team, ShotsOnTarget is the number of shots on target taken by a team, *Falls* is the number of falls committed per game, and Corners is the number of corners taken by a team. Linear regression analysis will be held for this research to examine the significance of those variables on the dependent one. The second dataset contains the final results of all the teams during the last 10 seasons, totalling to 200 observations. Some other broad factors are included in this dataset, e.g. *dribbles* per game, the *final position of the previous year*, etc. The main aim of the second dataset is to show some patterns and insights between variables via data visualization (i.e. graphs, plots, etc.)

Background Information

There are 20 clubs in the Premier League, which is the top division of Britain. Each of the teams plays 38 matches during the course of a season, playing the other clubs twice – once at home stadium, and once in the stadium of their opponents. For each win, a team receives 3 points, for a draw one point is given and no points are awarded for a loss. The final rankings of the teams are determined by their total points, then goal difference, and then goals scored. If still equal, teams are deemed to occupy the same position, however, if there is a tie for the championship, for relegation or for qualification to other competitions, a play-off match at a neutral venue decides rank. The top four teams of each season get their place in the next year's UEFA Champions League group stage. The fifth-placed team in the Premier League and the winner of the FA Cup qualify for the subsequent season's UEFA Europa League group stage. However, if the winner of the FA Cup also finished in the top five places in the Premier League or has won one of the UEFA's major tournaments, then the sixth-placed team qualifies for the UEFA Europa League group stage as well. Moreover, a system of promotion and relegation exists between the Premier League and the EFL Championship, the latter is the second division of Britain. Each year, the three lowest teams in the Premier League are relegated to the Championship, and the top two teams from the Championship get a promotion to the Premier League. And the third promoted team is being selected among the third, fourth, fifth and sixth-placed clubs via play-offs.

Research question

What factors have significant impact on the final position of a club in the Premier League

What was done on similar topic

https://www.researchgate.net/publication/316674581_Application_of_Multiple_Linear_Regression_Models_in_the_Iden

Application of Multiple Linear Regression Models in the Identification of Factors Affecting the Results of the Chelsea Football Team. By Margarita Castillo, Amelec Viloria, Heidi Posso, Alexander Elias Parody Muñoz

Analysis

Regreasion Results

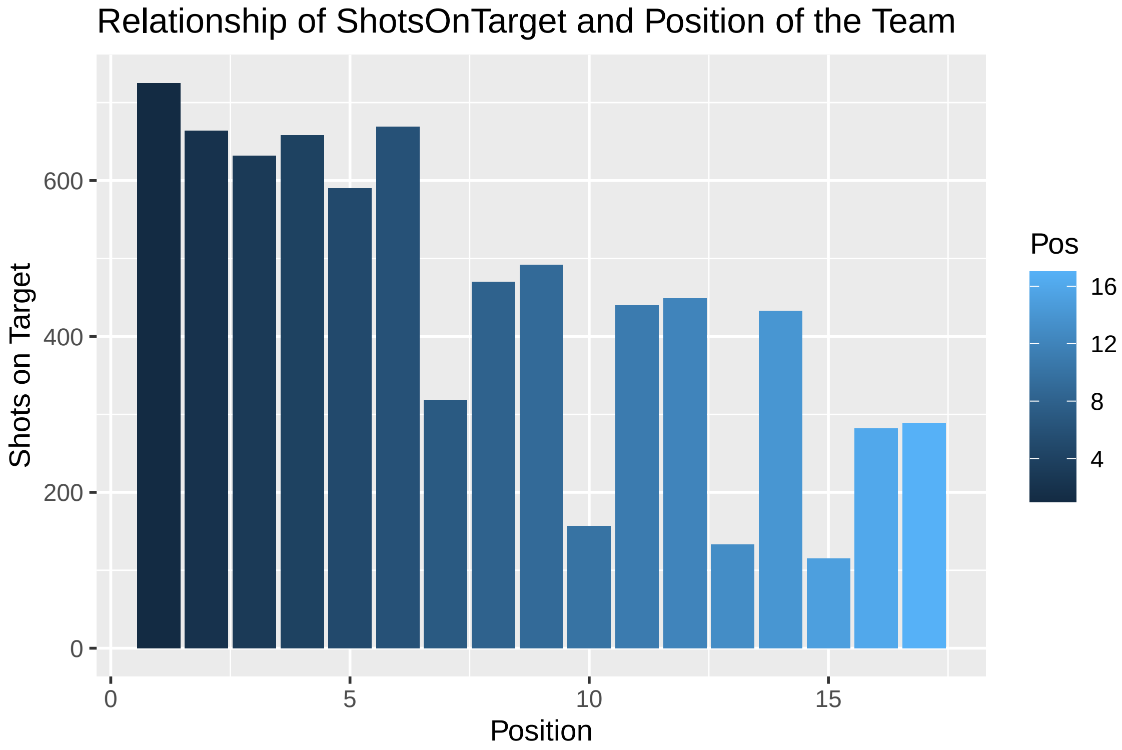
```
##
## Call:
## lm(formula = Pos ~ M_median + D_median + G_median + F_median +
##     Shots + ShotsOnTarget + Falls + Corners, data = Df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3113 -1.3023  0.4942  1.7853  4.1532
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -122.94091   89.66941  -1.371   0.1796
## M_median         0.19131    0.36121   0.530   0.5999
## D_median         0.35882    0.18602   1.929   0.0624 .
## G_median         0.05441    0.11970   0.455   0.6524
## F_median         0.15809    0.09638   1.640   0.1105
## Shots           -0.01999    0.01608  -1.243   0.2227
## ShotsOnTarget   -0.05766    0.03297  -1.749   0.0896 .
## Falls            0.01431    0.01147   1.247   0.2210
## Corners          0.02922    0.02617   1.116   0.2723
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.67 on 33 degrees of freedom
## Multiple R-squared:  0.758,  Adjusted R-squared:  0.6993
## F-statistic: 12.92 on 8 and 33 DF,  p-value: 3.3e-08
```

According to the results of the regression, only two of our independent variables turned out to be statistically significant on 10% confidence interval. One of them is the variable that indicates total shots on target by a particular team in a given season. As it could be intuitively anticipated, the shots on target have positive impact on the final position of a team, as the coefficient of the variable is negative. Our data, particularly, shows that one shot on target, on average, improves the position of a club by 0.058 units, all other things equal. The other variable that is statistically significant on 10% confidence interval is the defenders’ median height of a club. However, the coefficient of this variable seems to be a little bit surprising, as it is positive. Our regression results indicate that one more unit of this variable, on average, decreases the clubs position by 0.359, all other things equal. The results, however, tend to have low accuracy because of data availability issues and the topic, certainly, needs to be researched further in a more detailed manner.

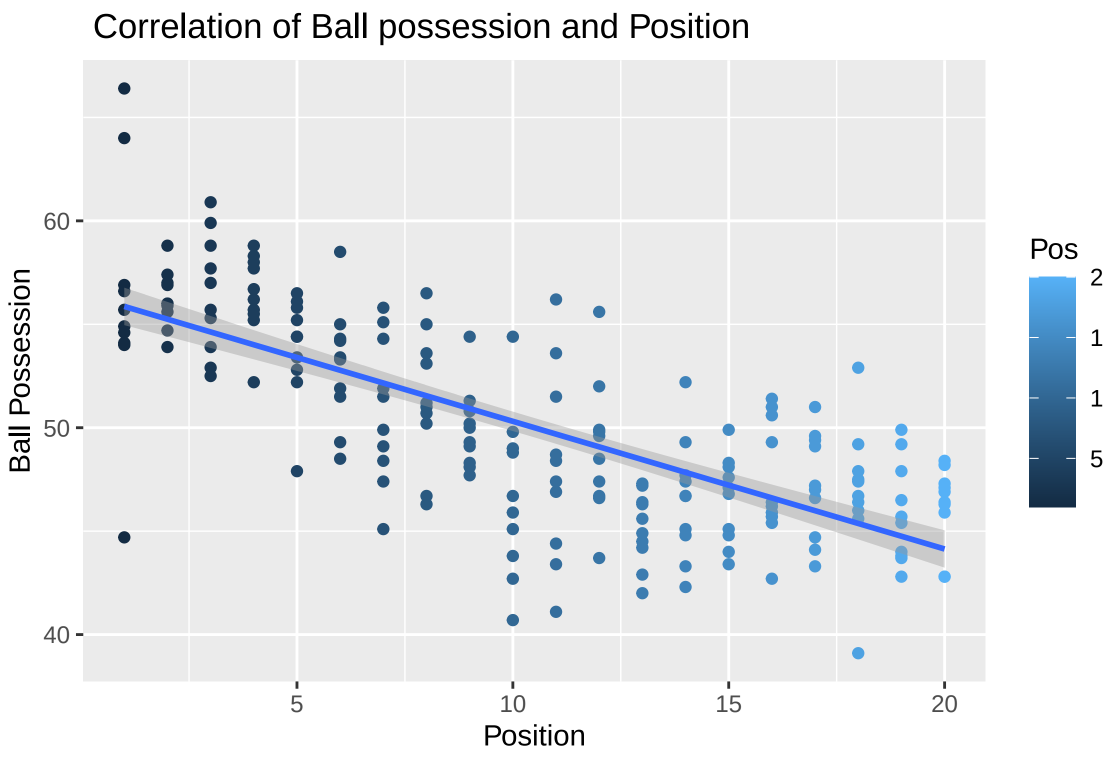
```
##
## Call:
## lm(formula = Pos ~ Shots + ShotsOnTarget + Falls + Corners, data = Df1)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3039 -1.2771  0.1389  1.2844  6.0840
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.45222    5.68201   3.423  0.00153 **
## Shots         -0.01072    0.01712  -0.626  0.53506
## ShotsOnTarget -0.07676    0.03480  -2.206  0.03369 *
## Falls          0.01375    0.01014   1.356  0.18321
## Corners        0.01174    0.02807   0.418  0.67811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.942 on 37 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.635
## F-statistic: 18.83 on 4 and 37 DF,  p-value: 1.606e-08
```

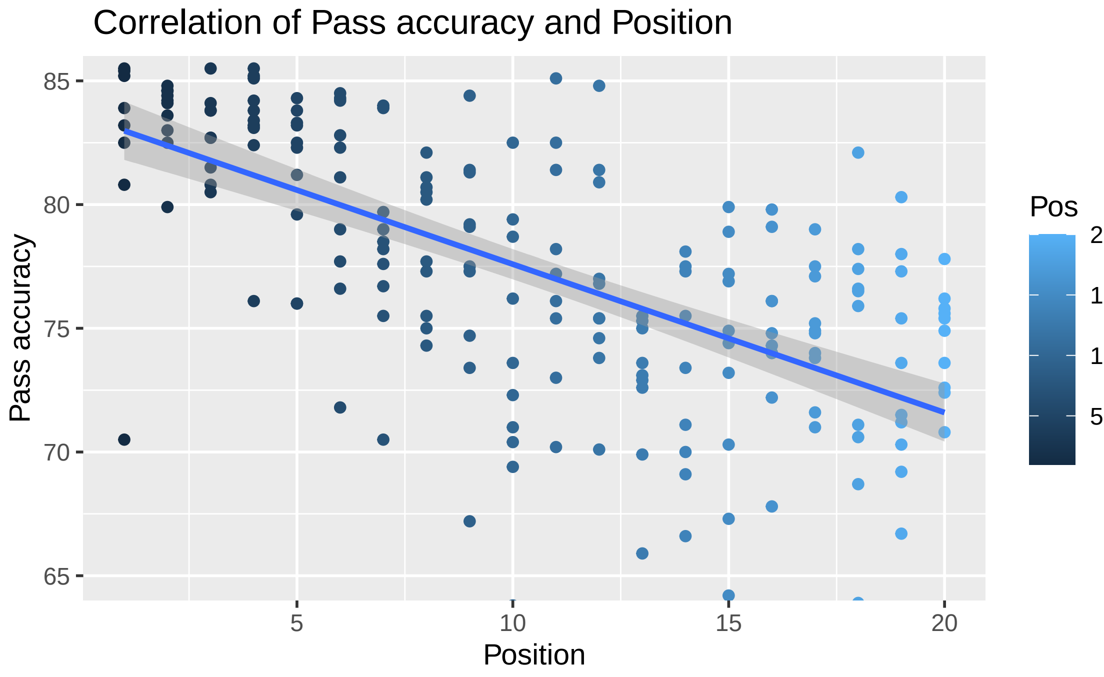
When we exclude the variables regarding players' heights, we get a result which is similar to the result of previous model, but the shots on target variable is statistically significant even on 5% confidence interval. The data shows that one unit increase of shots on target has a positive impact of 0.076 units on the final position of a club. Here the absolute value of shots on targets' coefficient is bigger than in the previous model. Again this model as well, may have some biases and accuracy issues as mentioned in further graph descriptions.



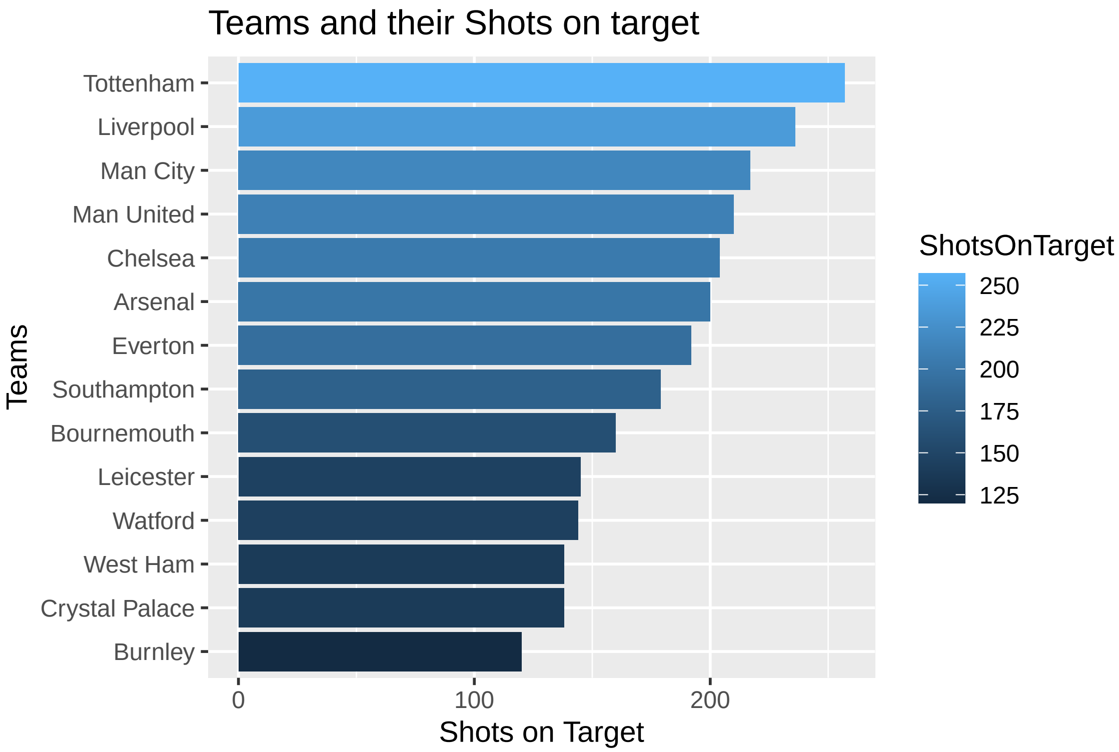
As can be noticed from the graph teams which had higher positions in the table made much more shots on target than those in lower positions, but it is not in the decending order as we see that in lower positions there were teams that also had comperatively high rate of shots on target.



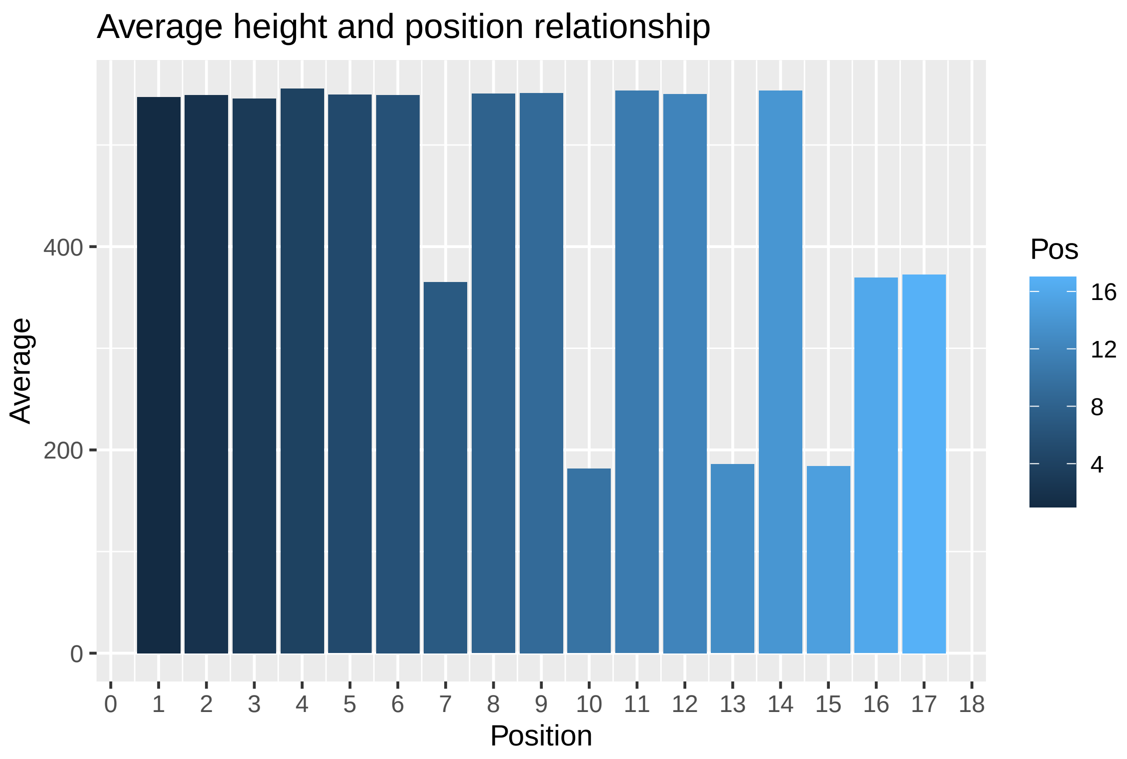
Here we want to see the relationship between the Position of the team and ball possession. Examining the second dataset we see that they are highly correlated. Which means that the teams which had higher positions in the table, also had higher ball possession in games. We see several outliars in the graph but the rest follow the trend. The blue line is the Regression line, also known as the best fit line or the least squares line. The shaded region near it shows our error rate. Because of structured data availability issue it was not feasible to include ball possession as an independent variable in the regression model, which of course had a negative impact on the accuracy of the model.



The graph above shows us the relationship of pass accuracy and Position of the team. Here the trend is not as smooth as in the previous graph, therefor we have bigger error rate than in previous one. Nevertheless we can conclude that the higher position the team had, the better was their pass accuracy in general. This term, which could have possibly increased the accuracy of the model, was also not included in the regression model due to the data availability problem.



This is the plot of 2018/19 Season's teams and their shots on target. In that season we see Tottenham, Liverpool, Manchester City and Manchester United are the four teams that had the highest number of Shots on target. Earlier we saw that ShotsOnTarget are correlated with the position of the team. In reality the table of first four places was like this (1.Man City, 2.Liverpool, 3.Chelsea, 4.Tottenham). We can notice that 3 out of 4 places that were in top 4 teams, also were in the top 4 teams that had the highest number of shots on target.



In the graph above we see the relationship of teams' average height calculated by taking the mean of the medians in each position (e.g for row 1 we got (M_median + D_median + G_median + F_median)/4). By this I wanted to see if there is a correlation between the final position of the team and the average height of the team. Here, by taking the data that we had, we cannot observe any correlation between those variables(there is no trend), but intuitively, it could happen that with the average height calculated in a different way, we would have a positive correlation of our variables.

Conclusion

So we can conclude that some of our variables such as shots on target, ball possession and pass accuracy have high correlation with the club final position, meantime by our regression models we rejected that some other features have a significant impact on the position. by graphs above we noticed some interesting patterns and visualizations that better explain regression result and give as a better perception.

Datasets

https://www.kaggle.com/datasets https://www.whoscored.com/ http://www.footballsquads.co.uk/eng/2018-2019/engprem.htm

Referances

https://www.premierleague.com/premier-league-explained?fbclid=IwAR0l33WauwhZDXSAQ73HsKfFk-a4x2mV9fqgc_ZrtsVQLreRHkAUuBgHnPU