zhao.3738_Project1_4194

Megan Zhao

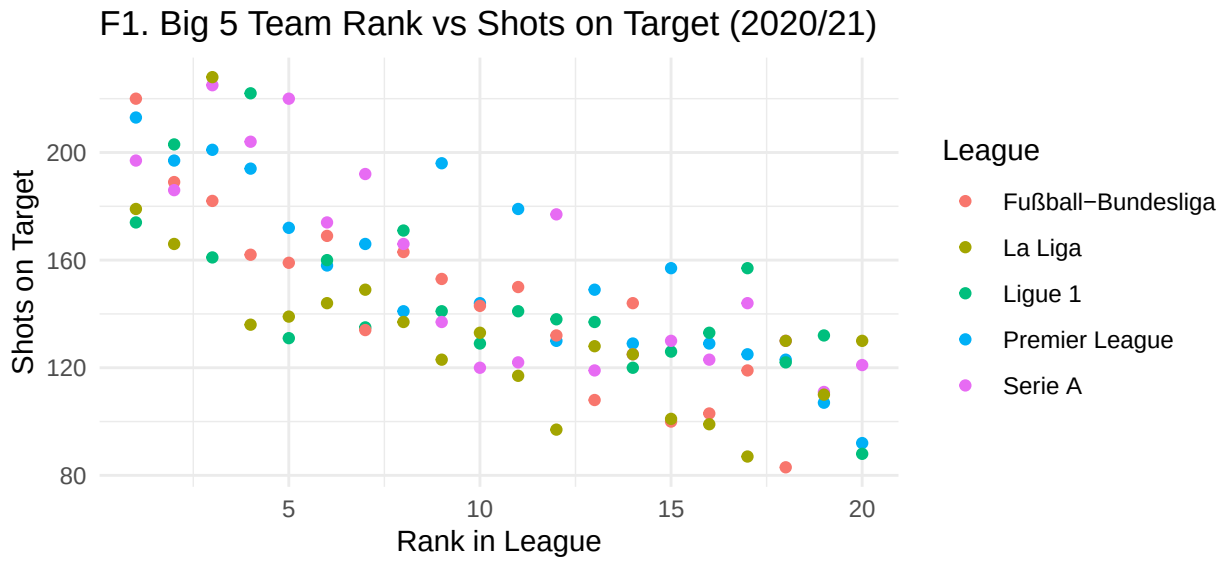
2025-02-16

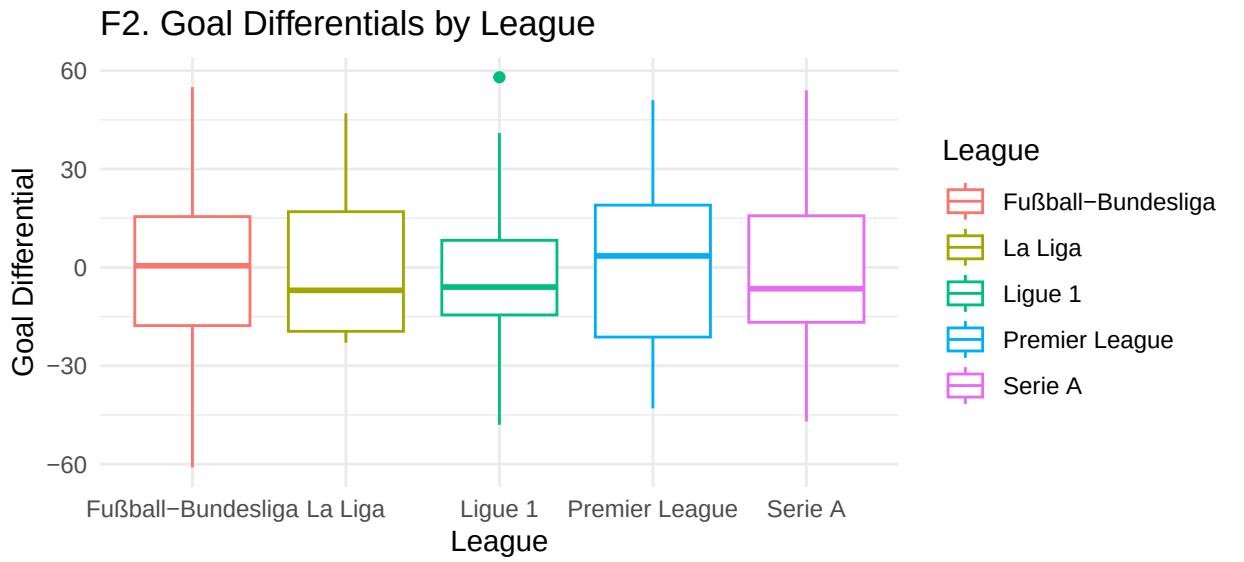
(a) Data Selection

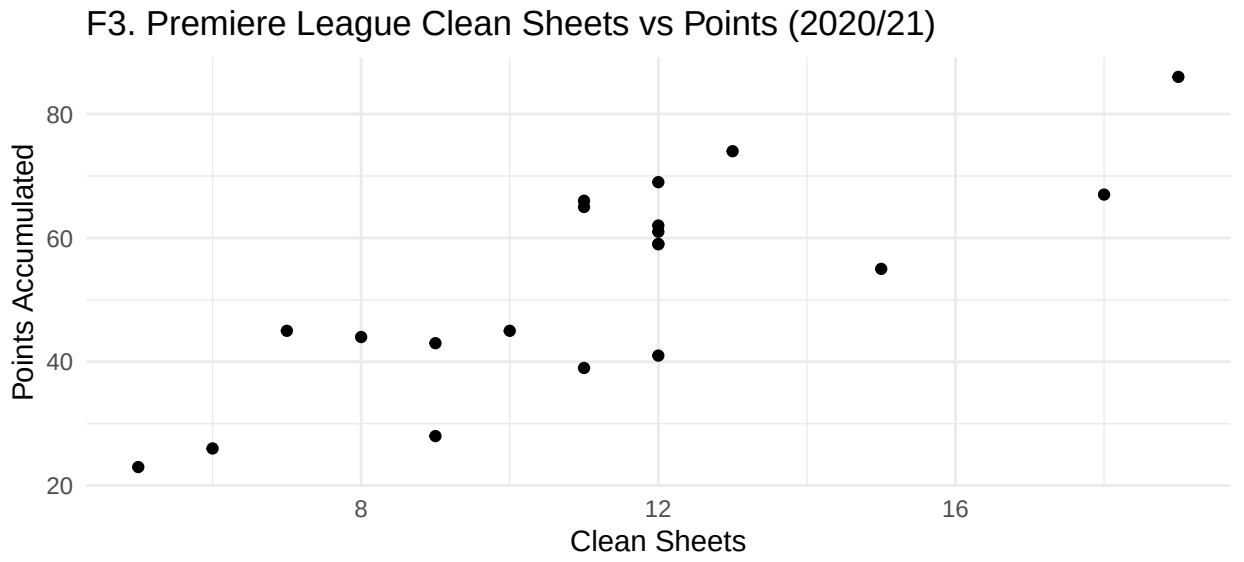
The dataset used in this project describes European men's football team statistics from the Big 5 European soccer leagues— Premier League (English), Ligue 1 (French), Bundesliga (German), Serie A (Italian) and La Liga (Spanish)— for the 2010/11 season up to 2020/2021 season. Each observation contains the name of the club, number of games played, number of wins, number of losses, number of draws, points accumulated (based on wins (+3) draws (+1) and losses (+0)) and other relevant football statistics such as players used, penalties, clean sheets, and average points per match. In these European soccer teams, season performance and overall European ranking (most look at UEFA coefficients) is based on each club's accumulated points over the current and recent seasons and performance in UEFA tournaments such as the Champion's League, the Europa League, and Conference leagues. The original data was taken from Kaggle.com with the original data source cited below.

Jordi Grau Escolano, & Albert Geli Taberner. (2021). Big 5 European football leagues: team and player stats [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5651722

(b) Exploratory Analysis







Based on the summary of the 2020-2021 season team data, no variables have missing data. In Figure 1 (F1), we can see a significant correlation between rank in league versus the number of shots on target the team had over the season. This is expected as each team's shots on target is likely correlated with their offensive dominance and goals scored, which contribute to wins, which contribute to ranking. In Figure 2, we can compare the total goal differential for teams (goals scored - goals conceded) across leagues. The goal differentials across the 5 leagues is not significantly different; however, we can interpret the plot to discern general league competitiveness or even-ness of matchups. For the Premiere League, which in the past couple of years has been widely considered as the most competitive European league, we see a clear positive median score differential, showing more teams are scoring more than they concede. Zero/Negative median score differentials can imply more competitive balance through the league, though overall wide variability in score differential seems to be quite similar between all leagues. In Ligue 1, an outlier indicates one team that performed significantly better than the rest. Looking at correlation between clean sheets (matches with 0 goals conceded) versus the total number of points for Premiere League teams (2020/21) in F3, we see a general positive correlation between points and clean sheets which can implicate a defensive performance's positive effect on overall team performance. Teams that have more clean sheets/defensive strength may secure more narrow wins or draws contributing to points. We see some variability which makes sense considering the role of offensive performance (goals scored) in points accumulated also.

(c) Model Building

```
soc2021 <- soc2021 %>%
  mutate(cards_earned = cards_yellow+cards_red)
m1 <- lm(goal_diff ~ shots_on_target + clean_sheets + cards_earned + League, data = soc2021)
summary(m1)
```

```
##
## Call:
## lm(formula = goal_diff ~ shots_on_target + clean_sheets + cards_earned +
##     League, data = soc2021)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.183  -8.932   1.651   7.061  23.078
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -95.741383   9.585914  -9.988 3.03e-16 ***
## shots_on_target        0.543553   0.043237  12.571  < 2e-16 ***
## clean_sheets           2.135941   0.376435   5.674 1.68e-07 ***
## cards_earned           0.003502   0.097442   0.036 0.971414
## LeagueLa Liga          0.445011   4.449090   0.100 0.920549
## LeagueLigue 1         -5.280034   3.918300  -1.348 0.181192
## LeaguePremier League -12.691586   3.719094  -3.413 0.000966 ***
## LeagueSerie A         -7.700816   4.280102  -1.799 0.075337 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.88 on 90 degrees of freedom
## Multiple R-squared:  0.8334, Adjusted R-squared:  0.8205
## F-statistic: 64.33 on 7 and 90 DF,  p-value: < 2.2e-16
```

```
m2 <- lm(goal_diff ~ shots_on_target*League + clean_sheets*League + cards_earned*League, data = soc2021)
summary(m2)
```

```
##
## Call:
## lm(formula = goal_diff ~ shots_on_target * League + clean_sheets *
##     League + cards_earned * League, data = soc2021)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.101  -6.764   1.072   6.609  24.656
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -95.525545  31.467632  -3.036  0.00326 **
## shots_on_target                      0.689795   0.086156   8.006 9.15e-12 ***
## LeagueLa Liga                       19.299375  37.771563   0.511  0.61083
## LeagueLigue 1                      -54.948474  37.738814  -1.456  0.14940
```

```
## LeaguePremier League                    6.687459  38.901976   0.172  0.86396
## LeagueSerie A                          33.307088  39.625950   0.841  0.40318
## clean_sheets                            0.550088   1.015163   0.542  0.58945
## cards_earned                           -0.139727   0.342049  -0.409  0.68403
## shots_on_target:LeagueLa Liga          -0.230927   0.131064  -1.762  0.08200 .
## shots_on_target:LeagueLigue 1          -0.134144   0.127135  -1.055  0.29462
## shots_on_target:LeaguePremier League   -0.257352   0.146846  -1.753  0.08361 .
## shots_on_target:LeagueSerie A          -0.254352   0.124722  -2.039  0.04480 *
## LeagueLa Liga:clean_sheets              1.685016   1.234185   1.365  0.17609
## LeagueLigue 1:clean_sheets              1.617649   1.200066   1.348  0.18157
## LeaguePremier League:clean_sheets       2.100995   1.538227   1.366  0.17591
## LeagueSerie A:clean_sheets              2.908409   1.458027   1.995  0.04956 *
## LeagueLa Liga:cards_earned              0.043153   0.376732   0.115  0.90910
## LeagueLigue 1:cards_earned              0.743961   0.399550   1.862  0.06637 .
## LeaguePremier League:cards_earned       0.004586   0.442238   0.010  0.99175
## LeagueSerie A:cards_earned             -0.267253   0.402834  -0.663  0.50901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.36 on 78 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8371
## F-statistic: 27.23 on 19 and 78 DF,  p-value: < 2.2e-16
```

Two distinct regression models used total shots on target, number of clean sheets, number of cards earned (both yellow and red cards), and League to predict a club's goal differential during the 2020/21 football season. Based on our knowledge of soccer and the exploratory analysis, many variables (ie goals and wins) are highly correlated, while others are connected (ie wins and points); therefore, these variables were chosen to avoid multicollinearity as independent predictors to one another. Shots on target indicates offensive dominance, while number of clean sheets focuses on a team's defending. Though teams that are better offensively may be strong all around and strong defensively, this is not necessarily the case. Slight correlation is not guaranteed. Cards earned and a team's league can also reasonably be defended to be independent of the other predictors.

In model one, which is a multiple linear regression, there is sufficient statistical evidence to conclude that shots on target (p >2e-16), number of clean sheets (p = 1.68e-07) and being in the Premier League compared to the baseline Bundesliga (p = 0.000966) effects the team's score differential (for each predictor holding all others constant).For each additional shot on target, For each additional shot on target, goal difference is expected to increase by 0.544. For each additional clean sheet, For each additional clean sheet, the goal difference is expected to increase by 2.136. We see little effect by a team's cards earned/discipline. Compared to the Bundesliga, a team playing in the Premier League is associated with a decrease in goal difference by -12.69. This implies that playing in the Premier League results in a significantly lower goal difference, though the median goal differential in the Premiere League was higher than the rest of the Big 5 in our exploratory analysis. Therefore, we will try a different model with interaction terms to allow the predictors to vary by league.

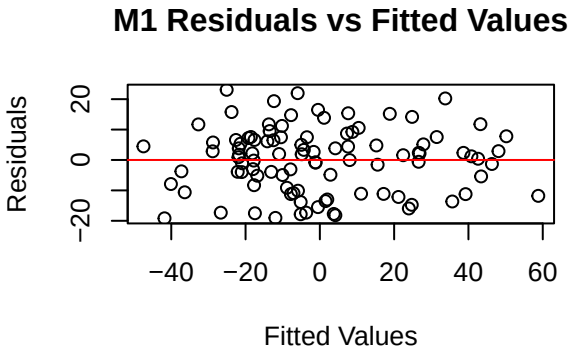
In model 2, we also add interaction effects between each of our continuous predictors and the team's league to the model. Still, we see a significant (p = 9.15e-12) effect by shots on target on the teams total goal differential, where an additional shot on target is associated with a 0.69 increase in goal differential. The impact of clean sheets and league is no longer significant. We can conclude that the impact of clean sheets might be conditional on the league a team plays in (in leagues where style of play is more focused on possession than aggressive strategies, clean sheets may be more common). This model implys that goal difference is complex and is likely influenced by factors that come with league-specific dynamics.

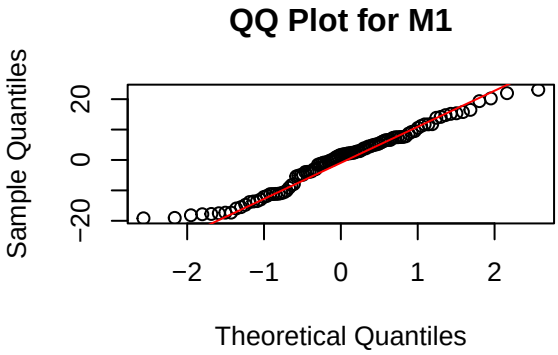
d) Model Evaluation

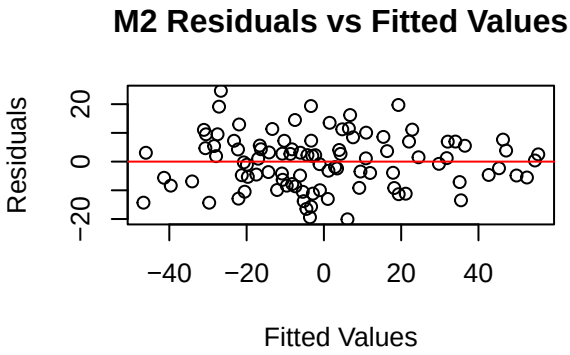
```
cor_matrix <- soc2021 %>%
  select(shots_on_target, clean_sheets, cards_earned) %>%
  cor()
cor_matrix
```

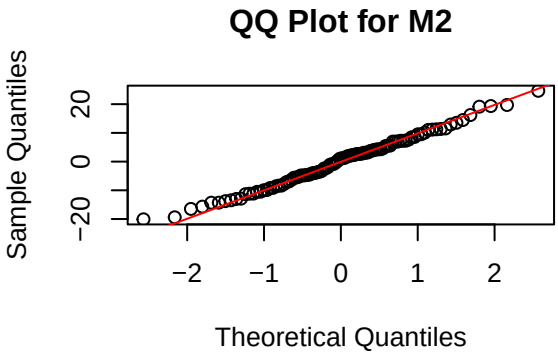
```
##                 shots_on_target clean_sheets cards_earned
## shots_on_target       1.0000000    0.5313181   -0.2854475
## clean_sheets          0.5313181    1.0000000   -0.1309110
## cards_earned         -0.2854475   -0.1309110    1.0000000
```

```
par(mfrow = c(2,2))
#m1
plot(fitted(m1), residuals(m1), xlab = "Fitted Values",
ylab = "Residuals", main = "M1 Residuals vs Fitted Values")
abline(h = 0, col = "red")

qqnorm(residuals(m1), main = "QQ Plot for M1")
qqline(residuals(m1), col = "red")

#m2
plot(fitted(m2), residuals(m2), xlab = "Fitted Values",
ylab = "Residuals", main = "M2 Residuals vs Fitted Values")
abline(h = 0, col = "red")

qqnorm(residuals(m2), main = "QQ Plot for M2")
qqline(residuals(m2), col = "red")
```










```
paste0("M1 BIC:", BIC(m1))
```

```
## [1] "M1 BIC:778.809015526663"
```

```
paste0("M2 BIC:", BIC(m2))
```

```
## [1] "M2 BIC:810.312755732436"
```

Firstly, based on the correlation matrix, none of the predictor variables chosen have a correlation higher than 0.7, so we can say that no two predictors are highly correlated and multicollinearity is not a major issue. Using multiple linear regression, we are assuming linear relationship between the dependent and independent variables, constant variance of errors (homoscedasticity), and normality and independence of residuals. Based on both Residuals vs Fitted Values plots for M1 and M2, the points appear randomly scattered above and below zero with no clear waving, funneling, or other patterns; this fulfills the linear, homoscedasticity, and indepence assumptions of multiple linear regression. In the M1 qqplot, we see a little bit of tailing off the diagonal qqline especially at the negative end, which implies that the residuals for M1 may not be truly normal and may have more extreme values. The heavy tail on the M2 qqplot is not as apparent, so M2 appears to better fulfills the normality assumption.

By comparing BICs of model 1 and 2, model 1 has the lower BIC (779<810) and I will therefore be selected as the better model as it better models fit and complexity is better balanced. This is not an unexpected outcome considering the number of parameters added to model 2 while considering interactions. The amount of interactive parameters did not significantly improve model fit. However, though we chose model 1 as the best model (a multiple linear regression model using shots on target, clean sheets, and league to predict a team's goal differential in teh 2020/21 season), based on our assumption assessments and experimental models, there are likely more complex interactions/predictors that contribute into each team's goal differential.

e) Analysis and Findings

The purpose of this analysis was to examine different factors that may influence the total goal differential between soccer teams in the Big 5 leagues during the 2020/21 football season. Goal differential can go beyond measuring a team's "success" (similar to ranking or wins) by more clearly depicting a team's dominance (goal differential being goals score minus goals conceded). We wanted to understand how league (which may have varying levels of competitiveness or play style, leading to closer games or overwhelmingly dominant teams, conversely), shots on target offensive authority), number of clean sheets (defensive authority), and cards earned (a more aggressive play style could increase goal differential or decrease it if penalties/opponent set pieces lead to more concessions) contributed to total goal differential. A multiple linear regression was conducted with the formula goal_diff ~ shots_on_target + clean_sheets + cards_earned. 98 observations were included.

At a significance level of 0.05, with all else being equal, teams in the Premier League have a significantly lower goal difference compared to the baseline league Bundesliga. All else equal as well, as shots on target increase, goal difference is expected to increase. This is intuitive as more shots on goal indicates that a team is creating more promising chances to score and perform better, driving up the score differential. All else equal there is a strong positive correlation between clean sheets and goal differential, which reinforces that strong defensive performances are associated with better overall dominance. (Specific coefficients are interpreted in part (c) Model Building above.) The number of cards earned had a coefficient of 0.004 ($p = 0.97$), so there is insufficient evidence of a statistically significant relationship between goal differential and the number of times a team was given a card as discipline. The Rsquared and F-statistic show that approximately 83% of variation in a team's goal differential could be explained by the independent variables in our model, and that our model generally significantly explains goal differential ($p = < 2.2e\text{-}16$).

Practically, it makes sense that teams that are overall dominant and have high score differentials (where they relatively score many goals and concede few) are both offensively and defensively strong. The Premiere League stands out in terms of goal differentials from the other leagues, possibly due to competitiveness, but the league a team plays in may interact with or influence the team's tactical style, opponents, etc which may help explain goal differential with more complexity that what is explained by our model. Based on the not significant impact of cards received, we can say that while discipline is important for maintaining team performance, stronger offensive and defensive strategies are more important, even if successful aggressive strategies lead to more penalties.