Project Part 3

Liam Quach, Dylan Li, and Brendan Callender

```
prem <- read_csv(here::here("data", "prem_multi_level.csv"))
```

```
Rows: 140 Columns: 20
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (3): Squad, Season, xG_cat
dbl (17): GF, GA, Pts, xG, xGA, Age, Poss, Sh, SoT, Dist, SoTA, CS, Expendit...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
prem_colors <- read_csv(here::here("data", "prem_team_colors.csv"), show_col_types = FALSE)
```

```
prem <- prem %>%
  mutate(xG_diff = GF - xG,
         xGA_diff = GA - xGA,
         SoT_diff = SoT - SoTA)
```

Part I: Proposal and Data Assembly

Research Questions

What are the most important factors associated with higher/lower point totals in the English Premier League between 2017 and 2023.

Is spending more money in the off-season associated with earning more points? Does this relationship change depending on how much the league as a whole spent?

Is having a better offense or defense more important for earning more points in a season?

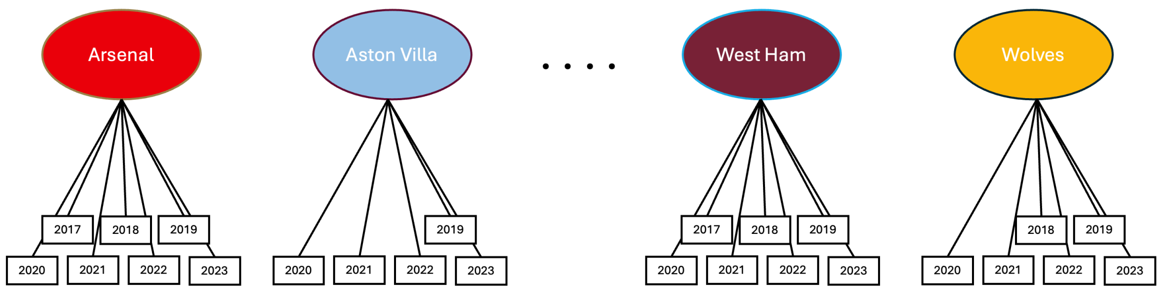
Does important is luck/chance with respect to premier league point totals. (We can measure luck as Expected goals scored vs Actual goals scored & Expected goals conceded vs Actual goals conceded)

Data

We are using data scraped from fbref and transfermark for the English premier league capturing the 2017 season up to and including the 2023 season. The response is the total points a team achieved for that given season. The predictors include variables relating to a teams offensive and defensive performance and a teams off-season expenditures.

Data Multi-level Structure

Figure: Multi-level Structure



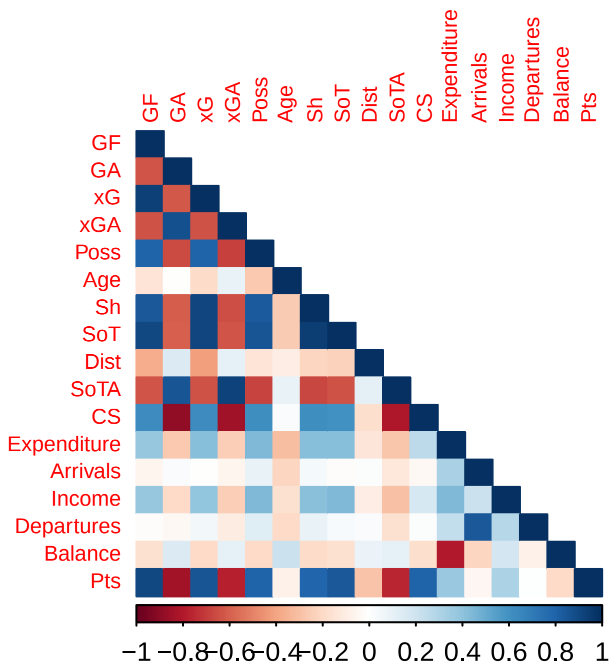
Variable Chart

Name	Role	Type	Values
Points	Response	Quantitative	>0
Goals/90	L1 Predictor	Quantitative	>0
Goals Against/90	L1 Predictor	Quantitative	>0
Net Spend	L1 Predictor	Quantitative	-inf, inf
Average Net Spend (for team)	L2 Predictor	Quantitative	-inf, inf
Luck Level	L1 Predictor	Categorical	(Lucky offense, Lucky defense), (Unlucky offense, Lucky defense), (Lucky offense, Lucky defense), (Unlucky offense, Unlucky defense)
...	...	...	...
Other ideas for L2			
Team Cateogry	L2 Predictor		Top 6, mid-table, relegation...

Name	Role	Type	Values

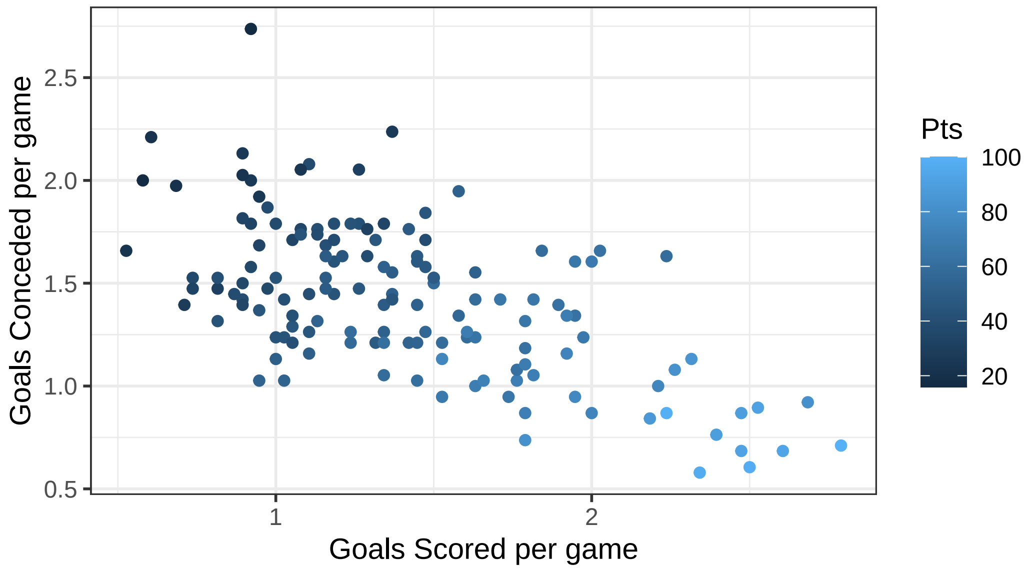
Part II: Exploratory Data Analysis

```
# Create a correlation plot to identify relationships between numeric variables
corr_matrix <- prem %>%
  select(GF, GA, xG, xGA, Poss, Age, Sh, SoT, Dist, SoTA, CS, Expenditure, Arrivals, Income,
  cor()
corrplot(corr_matrix, method = "color", type = "lower", tl.cex = 0.7, tl.pos = "lt")#, addCo
```



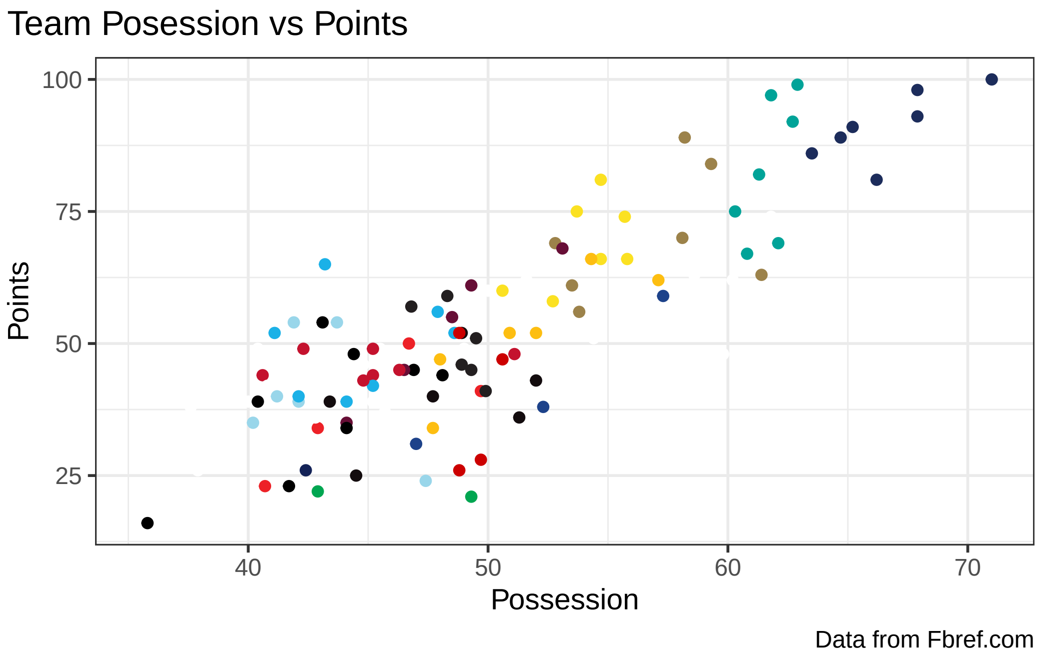

```
prem %>%
  ggplot() +
  geom_point(aes(x = GF, GA, color = Pts)) +
  theme_bw() +
  theme(plot.title.position = "plot") +
  labs(x = "Goals Scored per game",
       y = "Goals Conceded per game",
       legend = "Points",
       caption = "Data from Fbref.com",
       title = "Goals Scored and Goals Conceded vs Points")
```

Goals Scored and Goals Conceded vs Points

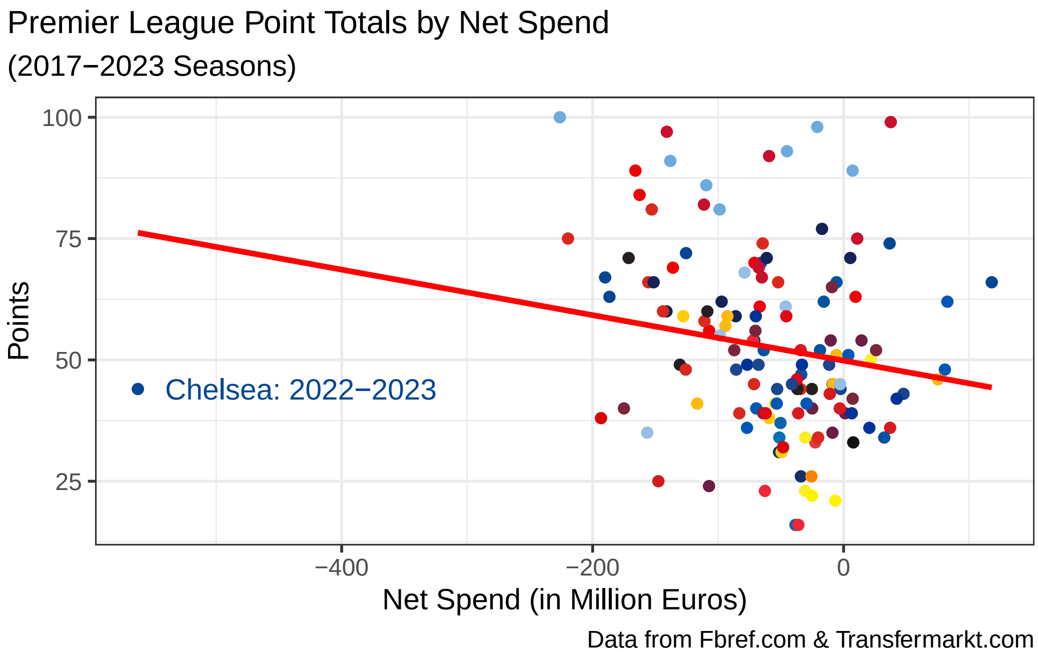


Data from Fbref.com

```
prem %>%
  left_join(prem_colors, by = "Squad") %>%
  ggplot() +
  geom_point(aes(x = Poss, y = Pts, fill = hex_fill, color = hex_color)) +
  theme_bw() +
  scale_fill_identity() +
  scale_color_identity() +
  theme(plot.title.position = "plot",
        legend.position = "none") +
  labs(y = "Points",
       x = "Possession",
       caption = "Data from Fbref.com",
       title = "Team Posession vs Points")
```

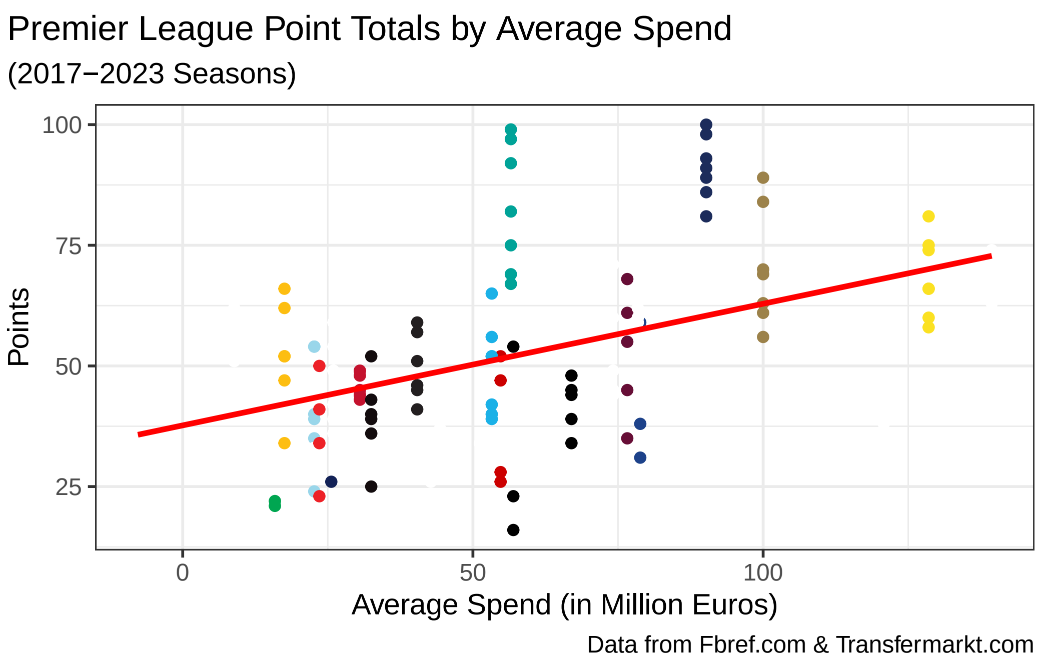



```r
# Scatter plot for net spend vs point total
prem %>%
  left_join(prem_colors, by = "Squad") %>%
  mutate(label = ifelse(Balance < -400, paste0(Squad, ": ", Season), "")) %>%
  ggplot(aes(x = Balance, y = Pts)) +
    geom_point(aes(color = hex_fill)) +
    geom_text(aes(x = Balance, y = Pts, label = label), hjust = -0.1, color = "#034694") +
    geom_smooth(method = "lm", se = FALSE, color = "red") +
    labs(title = "Premier League Point Totals by Net Spend",
         subtitle = "(2017-2023 Seasons)",
         caption = "Data from Fbref.com & Transfermarkt.com",
         x = "Net Spend (in Million Euros)",
         y = "Points") +
  theme_bw() +
  scale_color_identity() +
  theme(
    plot.title.position = "plot",
    plot.title = element_text(size = 12)
  )
```




```
prem %>%
  group_by(Squad) %>%
  summarize(Mean_Balance = -1*mean(Balance)) %>%
  left_join(prem %>% select(Squad, Pts), by = "Squad") %>%
  left_join(prem_colors, by = "Squad") %>%
  ggplot() +
  geom_point(aes(x = Mean_Balance, y = Pts, fill = hex_fill, color = hex_color)) +
  theme_bw() +
  scale_fill_identity() +
  scale_color_identity() +
  theme(plot.title.position = "plot") +
  labs(title = "Premier League Point Totals by Average Spend",
  subtitle = "(2017-2023 Seasons)",
       caption = "Data from Fbref.com & Transfermarkt.com",
       x = "Average Spend (in Million Euros)",
       y = "Points") +
  geom_smooth(aes(x = Mean_Balance, y = Pts), method = "lm", se = FALSE, color = "red")
```

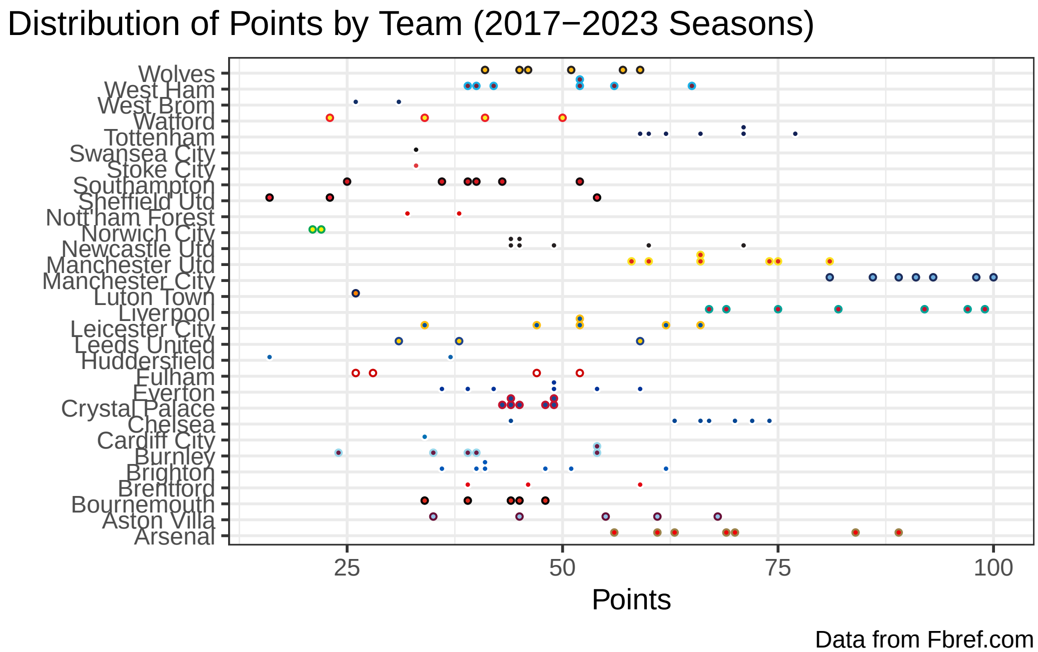
```
`geom_smooth()` using formula = 'y ~ x'
```



Part III: Modeling Results

Variability Across Teams (Level 2)

```
squad_colors <- read_csv(here::here("data", "prem_team_colors.csv"), show_col_types = FALSE)

prem %>%
  left_join(squad_colors, by = "Squad") %>%
  ggplot() +
  geom_dotplot(aes(x = Pts, y = Squad, fill = hex_fill, color = hex_color), binwidth = 1, do
  theme(legend.position = "none") +
  scale_fill_identity() +
  scale_color_identity() +
  theme_bw() +
  labs(
    x = "Points",
    y = "",
    title = "Distribution of Points by Team (2017-2023 Seasons)",
    caption = "Data from Fbref.com"
    ) +
  theme(plot.title.position = "plot")
```



Examining the graph of points across the seasons for the different teams, we can see a vast difference in both the ranges of points and variability in points depending on the team. This graph suggests strongly that the variation in points scored across the different teams is significant.

```
model00 <- lm(Pts ~ Squad, data = prem)
anova(model00)
```

```
Analysis of Variance Table

Response: Pts
           Df Sum Sq Mean Sq F value    Pr(>F)
Squad      29  37233 1283.89  12.848 < 2.2e-16 ***
Residuals 110  10992   99.93
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA with points and the teams grouping variable confirms what the graph has shown. There is a statistically significant variation in points across different teams.

Null Model

```
model0 <- lmer(Pts ~ 1 + (1 | Squad), data = prem)
summary(model0)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Pts ~ 1 + (1 | Squad)
   Data: prem

REML criterion at convergence: 1108.9

Scaled residuals:
     Min       1Q   Median       3Q      Max 
-2.02395 -0.60329 -0.08726  0.65915  2.11491 

Random effects:
 Groups   Name        Variance Std.Dev.
 Squad    (Intercept) 255.32   15.979  
 Residual              99.66    9.983  
Number of obs: 140, groups:  Squad, 30

Fixed effects:
            Estimate Std. Error t value
(Intercept)   47.388      3.091   15.33
```

The variance for the random effects of squad means measures how much variability there is in the average points depending on the team.

The variance for the residual measures the within team variability of points across the seasons. That is, when examining the same team, how much variance is there in points from season to season.

The intercept of the fixed effect is the least squared mean of points across seasons across teams.

```
performance::icc(model0)
```

```
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.719
  Unadjusted ICC: 0.719
```

The intraclass correlation of 0.713 means that the points across the seasons for each team is highly correlated. The ICC value is substantial, and it makes sense as a high performing team should score highly across different seasons, while weaker, lower performing teams would likely score low across different seasons.

Log Likelihood: -556.49 Deviance: 1112.97 AIC: 1118.97

Add Level 1 Vars

```r
model1 <- lmer(Pts ~ GF + GA + (1 | Squad), data = prem)
summary(model1)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Pts ~ GF + GA + (1 | Squad)
   Data: prem

REML criterion at convergence: 815.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.7861 -0.6433  0.0348  0.6699  3.2020

Random effects:
 Groups   Name        Variance Std.Dev.
 Squad    (Intercept)  0.9551  0.9773
 Residual             19.9221  4.4634
Number of obs: 140, groups:  Squad, 30

Fixed effects:
            Estimate Std. Error t value
(Intercept)   49.474      3.024   16.36
GF            24.099      1.042   23.14
GA           -21.903      1.325  -16.54

Correlation of Fixed Effects:
   (Intr) GF
GF -0.848
GA -0.909  0.582
```

```r
anova(model0, model1)
```

```
refitting model(s) with ML (instead of REML)

Data: prem
Models:
model0: Pts ~ 1 + (1 | Squad)
model1: Pts ~ GF + GA + (1 | Squad)
       npar     AIC     BIC  logLik deviance  Chisq Df Pr(>Chisq)    
model0    3 1118.97 1127.80 -556.49  1112.97                         
model1    5  829.01  843.72 -409.51   819.01 293.96  2  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model1 is significantly better compared to model0, we can see this based off the small p value from the chi square statistic. The chi squared statistic is found by taking the difference of the two loglik values and multiplying by 2, the df is the difference in number of parameters. The AIC value for model 1 of 829 is roughly 290 lower than the AIC of the null model.

```r
(255.32-0.9551)/255.32
```

```
[1] 0.9962592
```

99.6% of the variation in teams intercept is explained by including the goals scored (GF) and goals against (GA) variables. This means that GF and GA accounts for the majority of differences in points between teams, which is what we expected to see as teams who scored more goals and defended more goals will have more points.

```r
(99.66-19.9221)/99.66
```

```
[1] 0.8000993
```

80% of the season to season variation is explained by including the GF and GA variables.

Since both of these variables are significant, we will try adding in different 3rd variables to see how the model performs.

```r
model1_1 <- lmer(Pts ~ GF + GA + xG_cat + (1 | Squad), data = prem)
model1_2<- lmer(Pts ~ GF + GA + xG_diff + xGA_diff + (1 | Squad), data = prem)
model1_3<- lmer(Pts ~ GF + GA + Balance + (1 | Squad), data = prem)
model1_4<- lmer(Pts ~ GF + GA + SoT_diff + (1 | Squad), data = prem)
```

```r
anova(model1, model1_1)
```

```
refitting model(s) with ML (instead of REML)

Data: prem
Models:
model1: Pts ~ GF + GA + (1 | Squad)
model1_1: Pts ~ GF + GA + xG_cat + (1 | Squad)
         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model1      5 829.01 843.72 -409.51   819.01
model1_1    8 834.59 858.13 -409.30   818.59 0.4193  3     0.9362
```

```r
cat("\n\n")
```

```r
anova(model1, model1_2)
```

```
refitting model(s) with ML (instead of REML)

Data: prem
Models:
model1: Pts ~ GF + GA + (1 | Squad)
model1_2: Pts ~ GF + GA + xG_diff + xGA_diff + (1 | Squad)
         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model1      5 829.01 843.72 -409.51   819.01
model1_2    7 829.80 850.39 -407.90   815.80 3.2098  2     0.2009
```

```r
cat("\n\n")
```

```r
anova(model1, model1_3)
```

```
refitting model(s) with ML (instead of REML)

Data: prem
Models:
model1: Pts ~ GF + GA + (1 | Squad)
model1_3: Pts ~ GF + GA + Balance + (1 | Squad)
         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model1      5 829.01 843.72 -409.51   819.01
model1_3    6 830.49 848.14 -409.24   818.49 0.5227  1     0.4697
```

```r
cat("\n\n")
```

```r
anova(model1, model1_4)
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: prem
Models:
model1: Pts ~ GF + GA + (1 | Squad)
model1_4: Pts ~ GF + GA + SoT_diff + (1 | Squad)
         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model1      5 829.01 843.72 -409.51   819.01
model1_4    6 830.95 848.60 -409.48   818.95 0.0617  1     0.8039
```

None of these additional variables significantly improve our model as the chi-square value is small and the p value is large for all of these anova tests, so we will not be including any of these variables in our level 1 model.

Add Level 2 Vars

```r
prem <- prem %>%
  left_join(prem %>%
              group_by(Squad) %>%
              summarize(Balance_mean = -1*mean(Balance)),
            by = "Squad"
  )
```

```r
model2 <- lmer(Pts ~ GF + GA + Balance_mean + (1 | Squad), data = prem)
```

```
boundary (singular) fit: see help('isSingular')
```

```r
summary(model2)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Pts ~ GF + GA + Balance_mean + (1 | Squad)
   Data: prem
```

```
REML criterion at convergence: 815.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.8271 -0.6048  0.1171  0.6113  3.4867

Random effects:
 Groups   Name        Variance Std.Dev.
 Squad    (Intercept)  0.00    0.000
 Residual             19.96    4.467
Number of obs: 140, groups:  Squad, 30

Fixed effects:
              Estimate Std. Error t value
(Intercept)   49.11412    2.95766  16.606
GF            23.03892    1.05930  21.749
GA           -21.86604    1.29747 -16.853
Balance_mean   0.03120    0.01181   2.643

Correlation of Fixed Effects:
            (Intr) GF     GA
GF          -0.769
GA          -0.919  0.557
Balance_men -0.087 -0.362  0.056
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

```
anova(model1, model2)
```

```
refitting model(s) with ML (instead of REML)

Data: prem
Models:
model1: Pts ~ GF + GA + (1 | Squad)
model2: Pts ~ GF + GA + Balance_mean + (1 | Squad)
       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model1    5 829.01 843.72 -409.51   819.01
model2    6 824.35 842.00 -406.17   812.35 6.6622  1   0.009848 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model2 is significantly better compared to model1, we can see this based off the small p value of 0.0098 from the chi square statistic. The AIC value for model 2 of 824 is only roughly 5 lower than the AIC of model 1. This means that while adding in average spending of the team does improve the model fit, it is not as significant as model1 is to model0.

```
(0.9551-0)/0.9551
```

```
[1] 1
```

From model1 to model2, 100% of the variation in the intercepts of teams is explained, this means that along with goals and goals against, adding in average team spending perfectly explains all variation in the average points across the teams.

```
(19.9221-19.96)/19.9221
```

```
[1] -0.00190241
```

No within team variation in points is explained by average team spending, this makes sense as average team spending is a level 2 variable, so when examining an individual team, their average team spending will remain the same across all seasons.

No variables are insignificant so none will be removed.

Fit Random Slopes

```
model3 <- lmer(Pts ~ GF + GA + Balance_mean + (1 + GA | Squad), data = prem)
```

```
boundary (singular) fit: see help('isSingular')
```

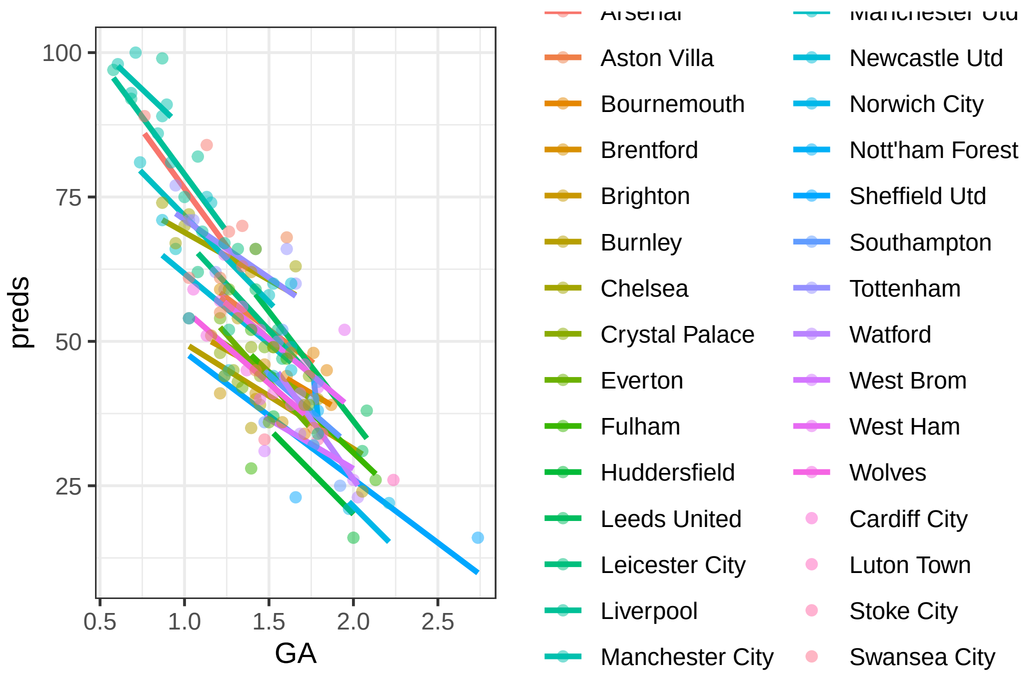
```
summary(model3)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Pts ~ GF + GA + Balance_mean + (1 + GA | Squad)
   Data: prem

REML criterion at convergence: 815.6

Scaled residuals:
```

```
    Min      1Q  Median      3Q     Max
-2.7386 -0.6304  0.0998  0.6088  3.3410

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Squad    (Intercept)  4.563   2.136
          GA           1.416   1.190    -1.00
 Residual             19.606   4.428
Number of obs: 140, groups:  Squad, 30

Fixed effects:
              Estimate Std. Error t value
(Intercept)   48.93810    3.00024  16.311
GF            23.06652    1.10339  20.905
GA           -21.81030    1.32018 -16.521
Balance_mean   0.03200    0.01234   2.593

Correlation of Fixed Effects:
            (Intr) GF     GA
GF          -0.750
GA          -0.913  0.518
Balance_men -0.078 -0.373  0.048
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

```r
teamAsFactor = factor(prem$Squad)
preds = predict(model3, newdata = prem)
ggplot(prem, aes(x = GA , y = preds , group = Squad, color = teamAsFactor )) +
geom_smooth(method = "lm", alpha = .5, se = FALSE) +
geom_point(data = prem, aes(y = Pts, color=teamAsFactor), alpha = .5) +
  theme_bw()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



We can see that as the goals against the team in that season increases, the points for that team in that season decreases. It also appears that the decrease in points for the goals scored against the team is greater if the team has higher scores.

The standard deviation of the slopes is 1.190, this means that all of the slopes are negative as the intercept for slopes is -21. And this also means that the slope generally does not change that much as the standard deviation of the slope is small relative to the slope fixed effect.

```
anova(model2, model3)
```

```
refitting model(s) with ML (instead of REML)

Data: prem
Models:
model2: Pts ~ GF + GA + Balance_mean + (1 | Squad)
model3: Pts ~ GF + GA + Balance_mean + (1 + GA | Squad)
       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model2    6 824.35 842.00 -406.17   812.35
model3    8 828.30 851.83 -406.15   812.30 0.0507  2      0.975
```

The difference in parameters between the two models is the variability in random slopes and the correlation of the slopes to the intercepts. Adding random slopes does not improve the model fit since the p-value is very large.

Cross Level Interaction

```r
model4 <- lmer(Pts ~ GF + GA + Balance_mean + GA*Balance_mean + (1 + GA | Squad), data = prem
```

```
boundary (singular) fit: see help('isSingular')
```

```r
summary(model4)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Pts ~ GF + GA + Balance_mean + GA * Balance_mean + (1 + GA |
    Squad)
   Data: prem

REML criterion at convergence: 820

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.8754 -0.6000  0.0706  0.6114  3.3230

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Squad    (Intercept)  3.498   1.870
          GA           1.035   1.017    -1.00
 Residual             19.722   4.441
Number of obs: 140, groups:  Squad, 30

Fixed effects:
                  Estimate Std. Error t value
(Intercept)      51.183854   4.208207  12.163
GF               23.107262   1.099031  21.025
GA              -23.393495   2.473397  -9.458
Balance_mean     -0.003402   0.048324  -0.070
GA:Balance_mean   0.025593   0.033953   0.754

Correlation of Fixed Effects:
            (Intr) GF     GA     Blnc_m
GF          -0.501
GA          -0.941  0.237
Balance_men -0.694 -0.142  0.826
GA:Balnc_mn  0.702  0.049 -0.847 -0.967
```

```
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

The interaction of 0.0256 means that as the average spending for a team increases, the effect of having goals scored against that team lowers the points for that team less.

```r
(1.416-1.035)/1.416
```

```
[1] 0.2690678
```

About 27% of the variation in slopes from model3 to model4 is explained by the cross level interaction, it is not much but some amount of variation in the slope is explained.

```r
anova(model3, model4)
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: prem
Models:
model3: Pts ~ GF + GA + Balance_mean + (1 + GA | Squad)
model4: Pts ~ GF + GA + Balance_mean + GA * Balance_mean + (1 + GA | Squad)
       npar   AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model3    8 828.3 851.83 -406.15    812.3
model4    9 829.6 856.08 -405.80    811.6 0.6959  1     0.4042
```

The cross level interaction model is not a significant model compared to the random slopes model.

Trying Longitudinal Models

```r
model0 <- lmer(Pts ~ 1 + Season + (1 | Squad), data = prem)
summary(model0)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Pts ~ 1 + Season + (1 | Squad)
   Data: prem

REML criterion at convergence: 1084.8
```

```
Scaled residuals:
    Min      1Q  Median      3Q     Max 
-1.8427 -0.6395 -0.1172  0.5967  2.1021 

Random effects:
 Groups   Name        Variance Std.Dev.
 Squad    (Intercept) 258.5    16.08   
 Residual             104.0    10.20   
Number of obs: 140, groups:  Squad, 30

Fixed effects:
                Estimate Std. Error t value
(Intercept)      48.3860     3.7996  12.734
Season2018-2019   0.3077     3.3531   0.092
Season2019-2020  -1.0911     3.3739  -0.323
Season2020-2021  -1.0521     3.3785  -0.311
Season2021-2022  -1.7493     3.3960  -0.515
Season2022-2023  -2.4091     3.4173  -0.705
Season2023-2024  -1.2154     3.4565  -0.352

Correlation of Fixed Effects:
            (Intr) S2018- S2019- S2020- S2021- S2022-
Ss2018-2019 -0.439                                   
Ss2019-2020 -0.440  0.512                            
Ss2020-2021 -0.441  0.507  0.522                     
Ss2021-2022 -0.443  0.509  0.533  0.524              
Ss2022-2023 -0.446  0.512  0.521  0.528  0.532       
Ss2023-2024 -0.453  0.507  0.522  0.523  0.519  0.539
```