Final Report

Dylan Li, Liam Quach, Brendan Callender

I. Introduction

The English Premier League (EPL) is the top tier of professional football (soccer) in England and is considered one of the most popular and competitive leagues in the world. The league is made up of twenty clubs (teams) that compete over a season for the Premier League title with new clubs added each year via a system of promotion and relegation. Each year, three new clubs are promoted from the second division based on the previous year's results with these promoted teams replacing the bottom three teams from the previous year's Premier League season.

Over the course of a season, each team plays a total of 38 matches, facing every other team twice—once at home and once away. Teams are rewarded points from each game as follows: 3 points for a win, 1 points for a draw, and 0 points for a loss. The team with the most points at the end of the 38-game season is crowned as the Premier League Champions.

add a little more maybe?

For our project, we are interested in exploring the following research questions:

1. What factors are associated with higher or lower point totals in the English Premier league?
2. Is spending more money in the off-season associated with earning more points the following season?
3. How do differences in expected goals scored vs actual goals scored and expected goals conceded vs actual goals conceded impact point totals?

II. Data Source & Methods

To answer our research questions, we collected English Premier League season-level data spanning from the 2017-2018 season up to the most recently completed 2023-2024 season. Data was collected from two sites: rbref.com and transfermarkt.com. The data collected from fbref includes performance related metrics for each team over the season as predictors as well as point totals for each team at the end of the season for our response variable. The performance metrics include total goals scored, total goals conceded, expected goals scored, expected goals conceded, average % possession, shooting metrics and more. The data collected from transfermarkt includes data relating to each teams expenditure and sales with respect to buying and selling players in the trasnfermarkt. This data includes money spent, money gained from sales, net spend, number of players bought, number of players sold and more. Money related variables are measured in thousands of euros.

Predictors relating to season totals such as goals scored and goals conceded were scaled down to per 90/ per game values for better interpretability. This was achieved by dividing these metrics by the total games played which is 38.

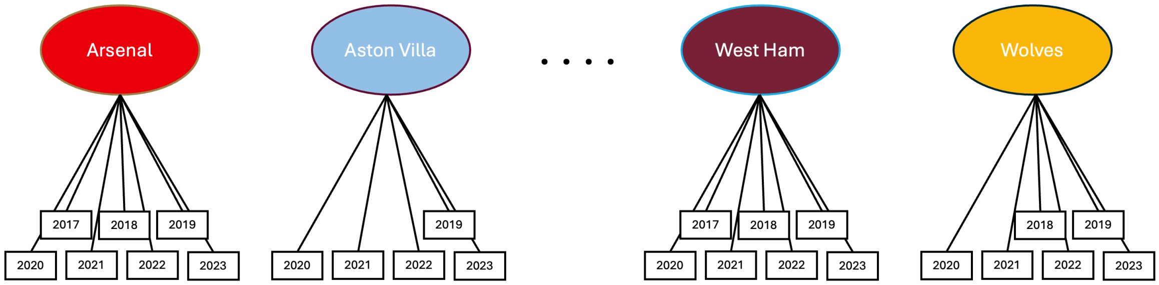
Name	Label	Role	Type	Values
Pts	Points	Response	Quantitative	>0
GF	Goals/90	L1 Predictor	Quantitative	>0
GA	Goals Against/90	L1 Predictor	Quantitative	>0
Balance	Net Spend	L1 Predictor	Quantitative	-inf, inf
Mean_Balance	Average Net Spend (for team)	L2 Predictor	Quantitative	-inf, inf
xG_cat	Actual vs Expected Metrics Category	L1 Predictor	Categorical	(Overperformed xG, Overperformed xGA), Underperformed xG, Overperformed xGA) Overperformed xG, Underperformed xGA) Underperformed xG, Underperformed xGA)
xG_diff	Actual vs Expected Goals Difference	L1 Predictor	Quantitative	
xGA_diff	Actual vs Expected Goals Against Difference	L1 Predictor	Quantitative	

See example rows of data below. (need to change which columns to show)

```
# A tibble: 3 x 11
  Club    Season    GF    GA    xG   xGA   Age  Poss Expenditure Income   Pts
  <chr>   <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>       <dbl>  <dbl> <dbl>
1 Chelsea 2017    1.63  1     1.43  0.89  26.7  55.6        260.   195.    70
2 Arsenal 2017    1.95  1.34  1.8   1.26  26.8  61.4        153.   162.    63
3 Everton 2017    1.16  1.53  1.07  1.38  26.7  45.5        203.   126.    49
```

To analyze the data, we will employ multi-level regression models, also known as hierarchical linear models. This approach is well-suited for the structure of the dataset, in which we have repeat observations for different clubs over several seasons. (See figure below)

Figure: Multi-level Structure



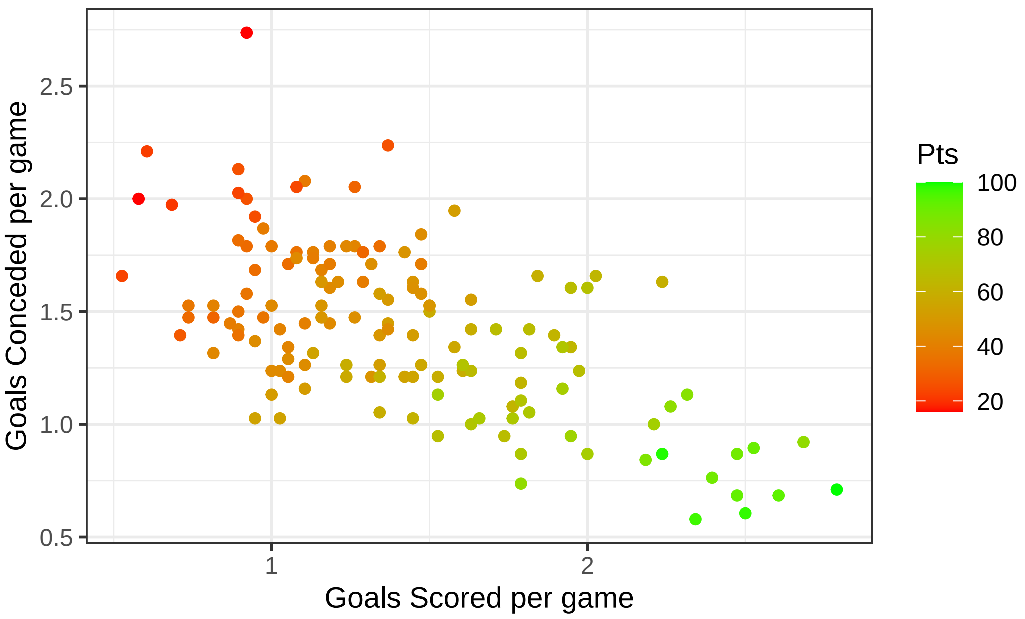
III. Results

Exploratory Data Analysis

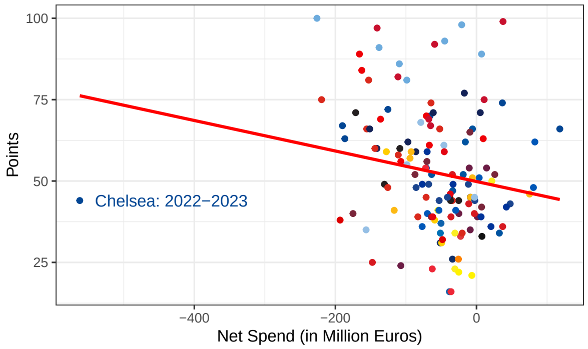
to do:

add writing

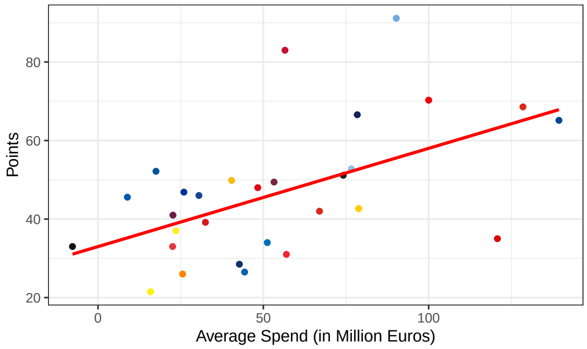
pick 1-2 more plots to show



Data from Fbref.com



Data from Fbref.com & Transfermarkt.com



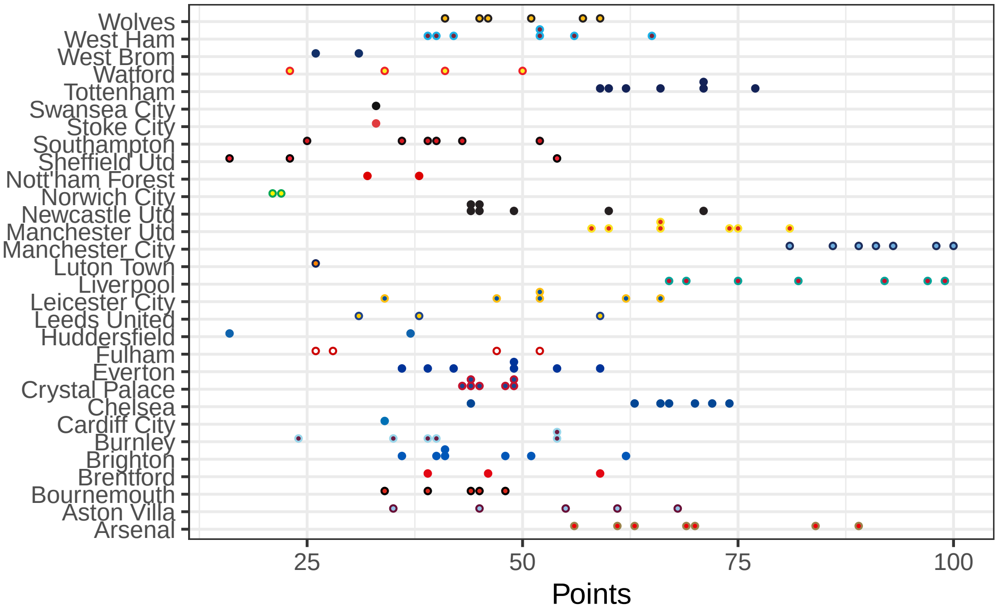
Data from Fbref.com & Transfermarkt.com

ANOVA

to do:

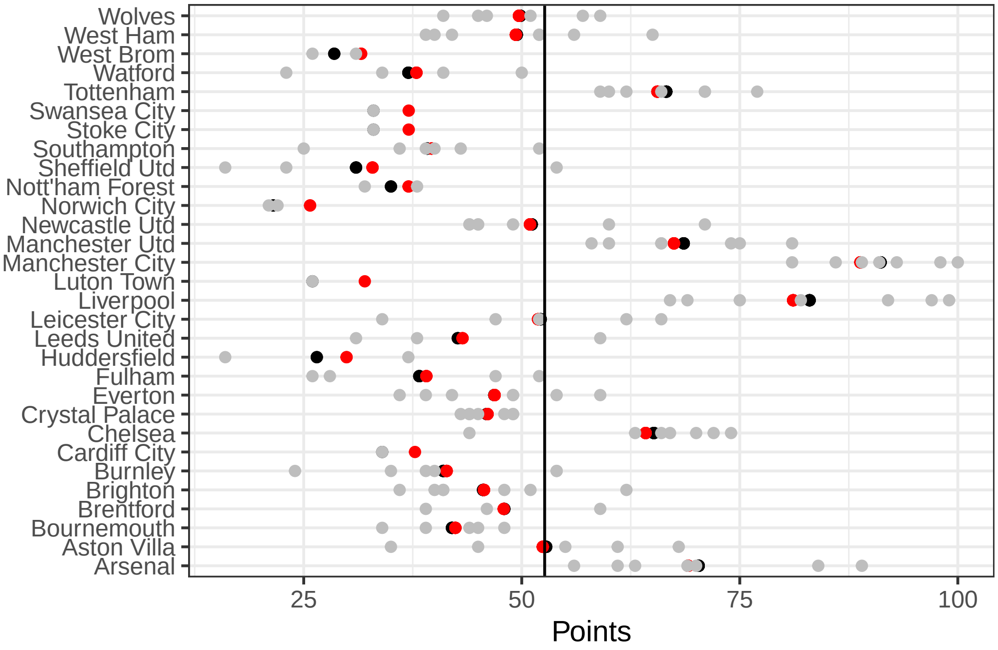
1. add writing

	df	SSE	MSE	F Statistic	P-Value
Club	29	37233	1283.89	12.848	< 0.0001
Residuals	110	10992	99.93		



Data from Fbref.com

Null Model



Model Fitting Process

add code here

write stuff here

Final Model

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

```
Linear mixed model fit by REML ['lmerMod']
Formula: Pts ~ GF + GA + Balance_mean + (1 | Club)
   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.
 Club     (Intercept)  0.00    0.000
 Residual             19.96    4.467
Number of obs: 140, groups:  Club, 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')
```

technical writing stuff here

V. Discussion

answer research questions

limitations

strengths and weaknesses

future steps

VI. Appendix

to do: add all model code stuff

add variable labels:

ANOVA

```
# A tibble: 2 x 6
  term         df  sumsq meansq statistic   p.value
  <chr>     <int>  <dbl>  <dbl>     <dbl>     <dbl>
1 Club         29 37233. 1284.       12.8  1.10e-23
2 Residuals   110 10992.   99.9      NA   NA
```