Cluserting Premier League and NFL Teams

Tom Grundy

09/03/2021

This project seeks to group Premier League and NFL teams based upon there attacking and defensive strengths. The analysis is performed in R with data obtained from https://www.football-data.co.uk/englandm.php and https://github.com/leesharpe/nfldata.

Project Aims

- Create an attacking and defensive metric for all Premier League and NFL teams
- Visualisation of the attacking and defensive metric to compare teams
- Cluster the teams based upon the attacking and defensive metric
- Create a final grouping of the teams to allow a user to find there favourite Premier League or NFL team and find similar teams in the alternative league.

Data Choices/Cleaning

First, we need to decide upon the data to be used in the project. Due to the Premier League having promotions and relegations, performing analysis over multiple seasons is difficult. Hence, we will use just the 2019/2020 season as this is the last fully completed season, which balances having enough data to create a relevant metric while still being relevant. As the 2020/2021 NFL season has finished we could use the most up to date season, however, in order to be consistent with the Premier League we also use the 2019/2020 season. This will also allow future analysis of the 2020/2021 season once the Premier League has finished.

Now, we need to decide upon an attacking a defensive metric for each team. We are not interested in the ability of each team as we could easily compare this across the leagues by looking at the teams standings. Hence, we will use the number of goals (points in the NFL) scored and conceded. While this will be correlated with the ability of the teams hopefully there should be more interesting conclusions than just matching the teams based upon there league placing.

Premier League Data

The Premier League data is available at https://www.football-data.co.uk/englandm.php in a .csv format. The spreadsheet contains all the results from the 2019/2020 Premier League season. From this we can clean the data to obtain the number of goals scored and conceded by each team over the season.

NFL Data

The NFL data was obtained from https://github.com/leesharpe/nfldata. This spreadsheet contains multiple summary statistics for each NFL team for each year. From this, we can clean the data so it shows the number of points scored and conceded by each team in the 2019/2020 season. We will abuse the NFL notation by calling renaming points as goals.

Table 1: Premier League Teams: Goals For (GF) and Goals Agaisnt (GA)

Team	GF	GA
Arsenal	56	48
Aston Villa	41	67
Bournemouth	40	65
Brighton	39	54
Burnley	43	50
Chelsea	69	54
Crystal Palace	31	50
Everton	44	56
Leicester	67	41
Liverpool	85	33
Man City	102	35
Man United	66	36
Newcastle	38	58
Norwich	26	75
Sheffield United	39	39
Southampton	51	60
Tottenham	61	47
Watford	36	64
West Ham	49	62
Wolves	51	40

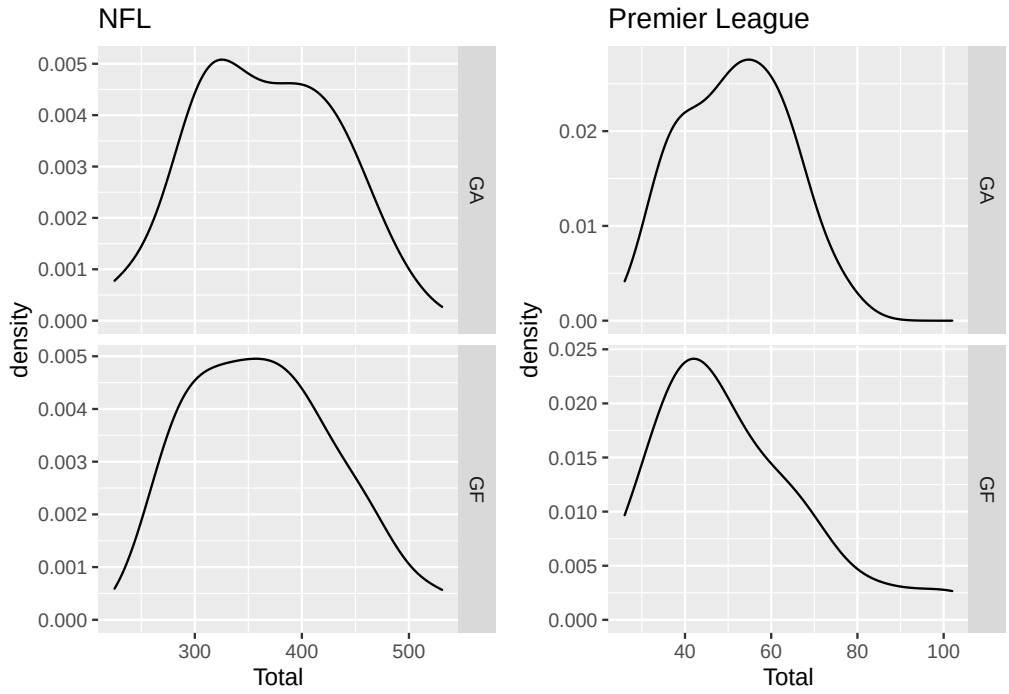
With the data cleaned and in a usable format we can now explore the data in more depth.

Data Exploration

First, we show density plots of the number of goals scored and conceded in each league using a Gaussian kernel. We show a density plot and not a histogram due to the small number of data points, which makes the histogram uninformative. Hence, we abuse the fact that the number of goals is discrete and plot an estimated density.

Table 2: NFL Teams: Points For (GF) and Points Agaisnt (GA)

Team	GF	GA
BUF	314	259
MIA	306	494
NE	420	225
NYJ	276	359
BAL	531	282
CIN	279	420
CLE	335	393
PIT	289	303
HOU	378	385
IND	361	373
JAX	300	397
TEN	402	331
DEN	282	316
KC	451	308
LAC	337	345
OAK	313	419
DAL	434	321
NYG	341	451
PHI	385	354
WAS	266	435
CHI	280	298
DET	341	423
GB	376	313
MIN	407	303
ATL	381	399
CAR	340	470
NO	458	341
TB	458	449
ARI	361	442
LA	394	364
SEA	405	398
SF	479	310



The density plot shows that the number of goals scored in the Premier League is right-skewed. This is probably because there is no maximum number of goals you can score but the least you can score is 0. Hence, the better teams can score more and the poor attacking teams are bounded below by the minimum they can score; 0. The NFL goals scored and conceded is less skewed. This is probably because the number of points/goals scored is much higher so the lower bound of 0 has less of an effect.

The number of goals scored in the Premier League is much lower than the number of points scored in the NFL. Therefore, to compare across the Premier League and NFL, we need to standardise the goals scored and conceded by each team. We do this by taking the number of goals scored/conceded by each team; subtracting the league mean; and dividing by the standard deviation in each league. This will produce a normalised attacking and defensive score that will be comparable across the two leagues.

The mean of the of the attacking and defensive score for the teams in each league will be 0. Hence, a positive attacking score means more goals scored than the average in the division and a negative attacking score means less goals scored than the average. We multiply the defensive scores by -1 so a large value indicates a better defensive performance; again the scores in each league will have a mean of 0.

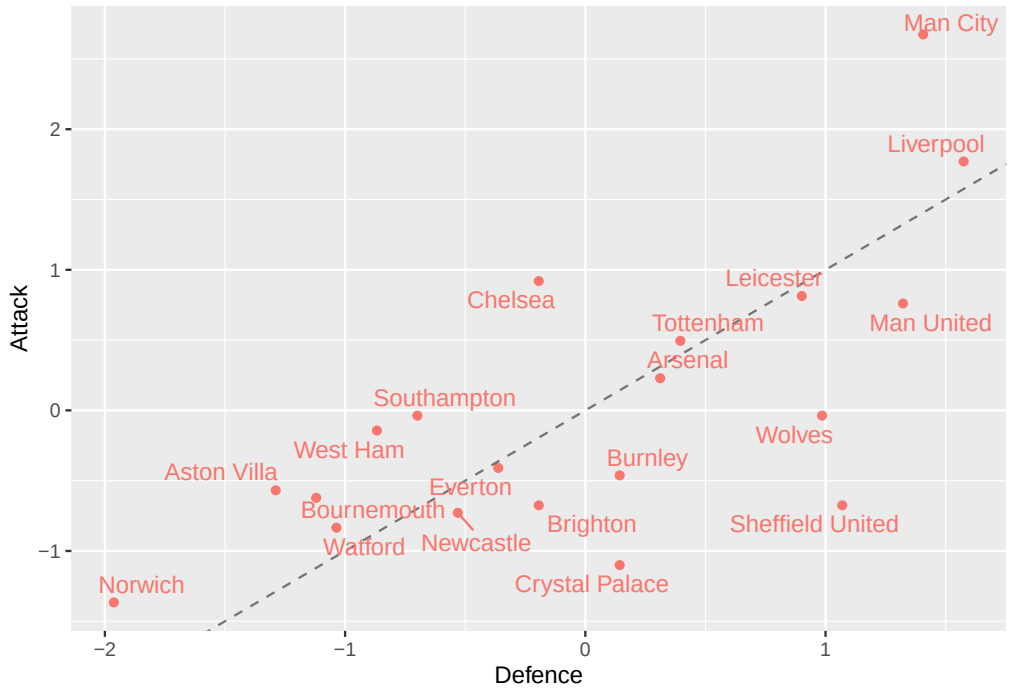
An alternative method of making the attacking scores comparable would be to minus the minimum number of goals scored/conceded and divide through by the range. However, the data is right skewed (the better teams score more and the minimum you can score is 0) and this effect was exaggerated when this type of scaling was applied.

Premier League

First, we explore the Premier League data. We can plot the relative attacking and defensive strengths of each team to get an initial look at the data.

Table 3: Attacking and Defensive Scores

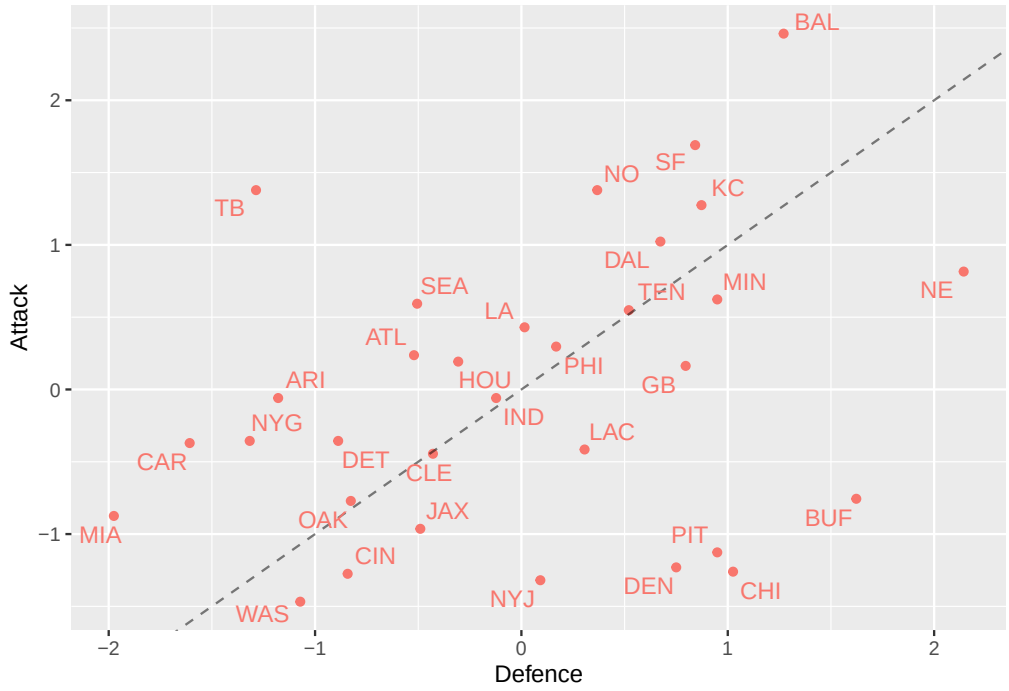
Team	Attack	Defence	League
BUF	-0.75603678	1.62272970	NFL
MIA	-0.87463079	-1.97483143	NFL
NE	0.81533379	2.14322791	NFL
NYJ	-1.31935831	0.09185262	NFL
BAL	2.46082561	1.27062797	NFL
CIN	-1.27488556	-0.84198239	NFL
CLE	-0.44472752	-0.42864558	NFL
PIT	-1.12664305	0.94914379	NFL
HOU	0.19271526	-0.30617542	NFL
IND	-0.05929700	-0.12247017	NFL
JAX	-0.96357629	-0.48988066	NFL
TEN	0.54849727	0.52049821	NFL
DEN	-1.23041280	0.75012977	NFL
KC	1.27488556	0.87259993	NFL
LAC	-0.41507902	0.30617542	NFL
OAK	-0.77086103	-0.82667362	NFL
DAL	1.02287330	0.67358591	NFL
NYG	-0.35578202	-1.31655429	NFL
PHI	0.29648501	0.16839648	NFL
WAS	-1.46760082	-1.07161395	NFL
CHI	-1.26006131	1.02568764	NFL
DET	-0.35578202	-0.88790870	NFL
GB	0.16306676	0.79605608	NFL
MIN	0.62261853	0.94914379	NFL
ATL	0.23718801	-0.52049821	NFL
CAR	-0.37060627	-1.60742093	NFL
NO	1.37865531	0.36741050	NFL
TB	1.37865531	-1.28593675	NFL
ARI	-0.05929700	-1.17877535	NFL
LA	0.42990327	0.01530877	NFL
SEA	0.59297003	-0.50518944	NFL
SF	1.68996457	0.84198239	NFL
Arsenal	0.22860701	0.31164275	Premier League
Aston Villa	-0.56885930	-1.28868487	Premier League
Bournemouth	-0.62202372	-1.12022933	Premier League
Brighton	-0.67518814	-0.19372387	Premier League
Burnley	-0.46253046	0.14318721	Premier League
Chelsea	0.91974448	-0.19372387	Premier League
Crystal Palace	-1.10050351	0.14318721	Premier League
Everton	-0.40936604	-0.36217941	Premier League
Leicester	0.81341564	0.90123713	Premier League
Liverpool	1.77037521	1.57505928	Premier League
Man City	2.67417036	1.40660374	Premier League
Man United	0.76025122	1.32237597	Premier League
Newcastle	-0.72835256	-0.53063494	Premier League
Norwich	-1.36632561	-1.96250702	Premier League
Sheffield United	-0.67518814	1.06969267	Premier League
Southampton	-0.03721509	-0.69909048	Premier League
Tottenham	0.49442911	0.39587051	Premier League
Watford	-0.83468140	-1.03600156	Premier League
West Ham	-0.14354394	-0.86754602	Premier League
Wolves	-0.03721509	0.98546490	Premier League



We can see there is clear positive correlation between the Attacking and Defensive scores as we would expect! The dashed line shows when the defensive score is equal to the attacking score (this is not the line of best fit!). We can see the teams above the line have a better attack than defence (Chelsea being the main example) while some have a better defence than attack (Sheffield United being the best example).

NFL

Now lets look at the NFL teams.



We see a similar positive correlation between attack and defence, however, this doesn't seem as strong as in

Table 4: Covariance of Premier League attack and defence scores

	Attack	Defence
Attack	1.0000000	0.6940305
Defence	0.6940305	1.0000000

Table 5: Covariance of NFL attack and defence scores

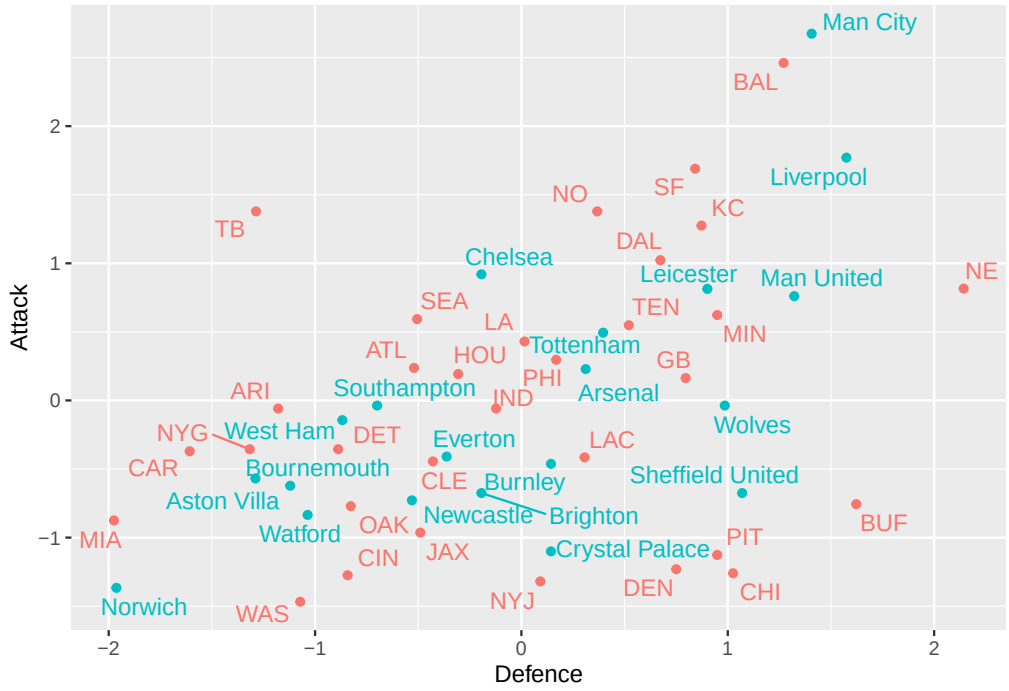
	Attack	Defence
Attack	1.0000000	0.3156018
Defence	0.3156018	1.0000000

the Premier League side.

This shows the NFL teams have less correlation between there attacking and defensive scores. As NFL teams have separate attacking and defensive teams we would expect less correlation between their attacking and defensive scores. Where as in the Premier League all players contribute to the attack and the defence.

Combination of Leagues

As the attacking and defensive scores of teams from both leagues have been scaled we can compare them directly. The plot below shows all the teams across the two leagues.



Here we can see the different covariances between the attack and defence scores for the different leagues. The NFL teams have less correlation and therefore we get more extreme examples of teams with a high attacking but low defensive score (Tampa Bay) and more examples of those with a higher defensive and low attacking score (Buffalo). This effect is exaggerated as there are more NFL teams than Premier League teams.

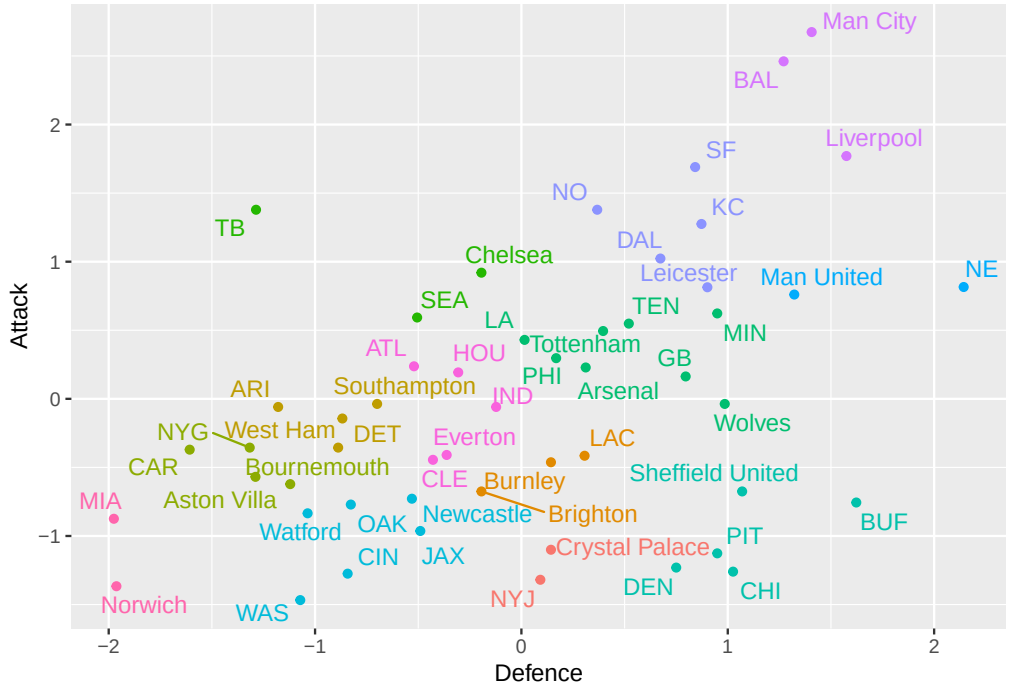
We could transform the scores from both leagues to remove the covariance. However, this will lead to uninterpretable axis on our graphics. As a result, we choose to stick with the original data to aid interpretability of the attacking and defensive scores and therefore the clustering of the teams.

Clustering

Now we can attempt to cluster points based upon their attacking and defensive strengths. We will use k-means clustering to achieve this.

The hardest part of k-means clustering (and many clustering algorithms) is deciding upon the number of clusters to use. Usually, we would use some diagnostic tool such as an elbow plot, however, we have a specific objective in mind: we want to group similar teams from the two different leagues. Therefore, it makes sense that we take the maximum number of clusters with the condition that there is at least one team from each league in each cluster.

The k-means algorithm doesn't always converge to a unique solution and hence can produce different clusters. This is especially true in our scenario where the data points are not well separated. Do to this non-unique convergence and using our rule for obtaining the number of clusters, we will get different numbers of clusters for different runs of the k-means algorithm. In an ideal scenario, we desire as close to a one-to-one matching of the teams as possible, hence the more clusters the better. As a result, we let the k-means algorithms run multiple times and select the run that returns the most clusters, essentially optimizing the k-means clustering for our objective - as many clusters as possible.



This gives us 13 clusters of Teams and visually they look reasonable. See the bottom of the page for the full list of groupings. We have some interesting groupings:

1. **Miami and Norwich:** True underdogs, both these teams have poor attacking and defensive strengths compared to the teams in their leagues. Interestingly, there were actually 5 worse teams in the NFL than Miami in terms of winning percentage. This shows grouping by attacking and defensive strength gives more insight than by purely using league position.

2. **Baltimore, Man City and Liverpool:** The top dogs, these teams were the best in the league for attacking and defensive strengths. This is to be expected with Man City and Liverpool well ahead in the premier league standings and the Baltimore Ravens were the number 1 team in the NFL regular season.

3. **Tampa Bay, Seattle and Chelsea:** All-out attack, these teams have significantly more success attacking than they do defensively. The 'we'll score more than you' sort of teams.

4. **Denver, Pittsburgh, Chicago, Buffalo and Sheffield United:** You shall not pass, these teams have rock solid defences although they don't offer much in the attacking sense.
5. **New England, Manchester United** The most hated teams? Arguably the two most hated teams in both leagues have been grouped together. This is more coincidence and because they are in the best-of-the-rest teams that favour defence over attack.

Summary and Potential Extensions

In this project, we have grouped together similar Premier League and NFL teams based upon their respective attacking and defensive strengths. We scaled the goals/points scored and conceded by teams in each league to put them on a comparative scale. We then used k-means clustering to cluster the teams to see identify similar teams across the leagues.

We could extend this work further by considering other metrics for each team in addition to attacking and defensive strength. These could include:

- Home strength - this may help separate teams that perform better at home than away and vice versa.
- Average attendance
- Number of 'star' players

These additional metrics could provide more separation between the teams and therefore provide a better clustering.

We could also look at alternative clustering techniques. One of the downfalls of k-means clustering is the solution is not unique and the algorithm can converge to different solutions. Here we optimized this to find a run which produced a high number of clusters. Alternatively, we could use a method such as Hierarchical clustering which converges to a unique solution.

Achievemnt of Aims

- Create an attacking and defensive metric for all Premier League and NFL teams
  - Scaled the number of goals scored and conceded by each team by the league mean and variance.
- Visualisation of the attacking and defensive metric to compare teams
  - Presented labelled scatter plots using `ggplot` to show each teams attacking and defensive strengths.
- Cluster the teams based upon the attacking and defensive metric.
  - Used k-means clustering to group the teams, optimizing to have a maximal number of clusters and at least one team from each league in each cluster.
- Create a final grouping of the teams to allow a user to find there favourite Premier League or NFL team and find similar teams in the alternative league.
  - Presented a table which shows which Premier League teams are grouped with which NFL teams.

Full Groupings

Cluster	Team
1	NYJ
	Crystal Palace
2	LAC
	Brighton
	Burnley
3	DET
	ARI
	Southampton
	West Ham
4	NYG
	CAR
	Aston Villa
	Bournemouth
5	TB
	SEA
	Chelsea
6	TEN
	PHI
	GB
	MIN
	LA
	Arsenal
	Tottenham
	Wolves
7	BUF
	PIT
	DEN
	CHI
	Sheffield United
8	CIN
	JAX
	OAK
	WAS
	Newcastle
	Watford
9	NE
	Man United
10	KC
	DAL
	NO
	SF
	Leicester
11	BAL
	Liverpool
	Man City
12	CLE
	HOU
	IND
	ATL
	Everton
13	MIA
	Norwich