**The English Premier League:**

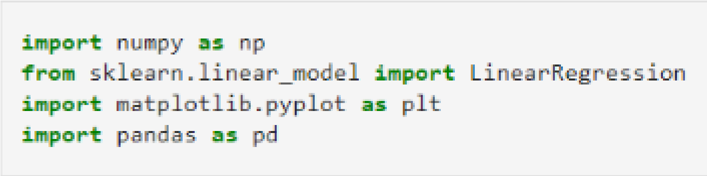
The English premier league, which is often referred to as the “EPL” is the top level of the English football league system. It consists of 20 teams who play against each other two times a year. A team plays thirty eight games in total and a single season consists of 760 games. The goal difference is a metric that is calculated by subtracting the number of goals conceded by a team from the number of goals scored. In general, the team with a larger goal difference will often win more games and finish higher in the table as winning a game results in 3 points. Similarly, a draw equals a point and a game lost is equal to 0 points.

Therefore, a linear relationship should exist between goal difference and the points earned by the team at the end of the season. With the help of linear regression, basic predictions can be made from the available data and number of deductions can be made by studying the linear relationship between the goal difference and points. The basics behind applying linear regression in football data, or any sports data in general is to understand how a quantity in an earlier period of time can help predict the same quantity during a later period.

In order to explore the relationship, we applied linear regression to the final standings of the English premier league seasons between 2010/11 and 2020/21. A table was created to record the final ranks, points, goal scored and the goals allowed . Goal difference was later calculated by subtracting the goals allowed by a team to the number of goals the team scored. Using various machine learning and mathematical libraries in python, we were able to analyse the impact of goal difference on the final points at the end of the season.

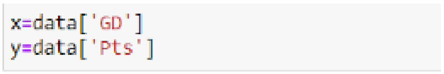
The following python libraries were used for the development of this project:

- pandas: fundamental package for data analysis and manipulation.
- numpy : mathematical functions and support to operate on multidimensional arrays and matrices.
- sickit-learn: machine learning library that features regression algorithm and allows integration with matplotlib for plotting.
- matplotlib: plotting library for python.

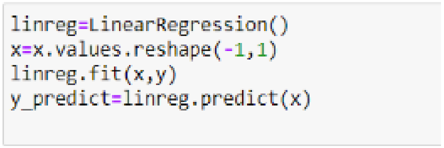


```
import numpy as np
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
import pandas as pd
```

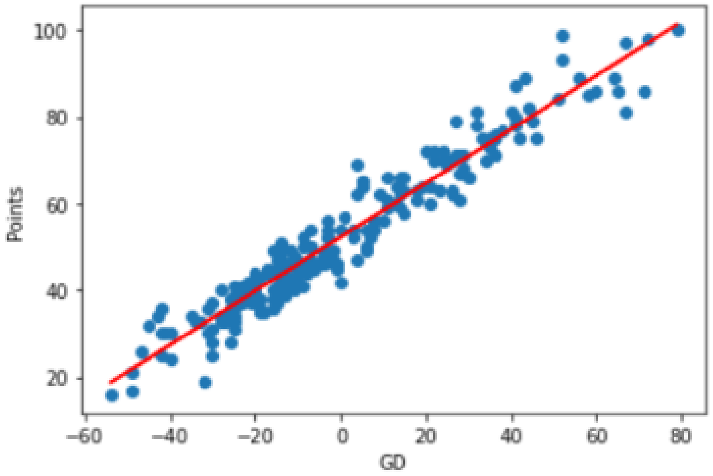
X-axis, the independent variable was the goal difference and the dependent variable, Y-axis was the total number of points earned at the end of the season. Each data point was imported to the plot as well as the best fit line was generated. The printed plot is shown below. From the plot, it was observed that a linear relationship existed between the two given variables.



```
x=data['GD']
y=data['Pts']
```

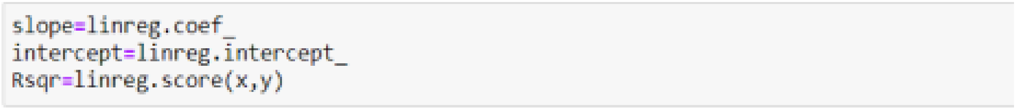


```
linreg=LinearRegression()
x=x.values.reshape(-1,1)
linreg.fit(x,y)
y_predict=linreg.predict(x)
```

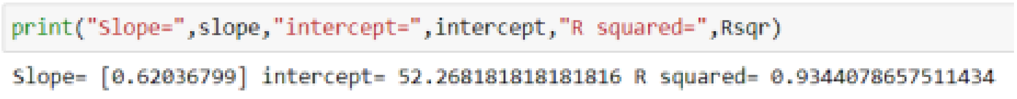


**Goal Difference Vs Points**

In the plot, the line of best fit has the α equal to the gradient. Similarly, the β is equal to the y-intercept. The X-axis gives the goal difference and the Y-axis gives the final season points of the clubs. From the calculated gradient, we can infer that if a club's goal differential increased by one, then at the end of the season the total points of that club will increase by 0.62. Similarly, from the y-intercept, we can infer that if a club has the same amount of goals that it concedes then it will finish the season with a total of 52.26 points.



```
slope=linreg.coef_
intercept=linreg.intercept_
Rsqr=linreg.score(x,y)
```



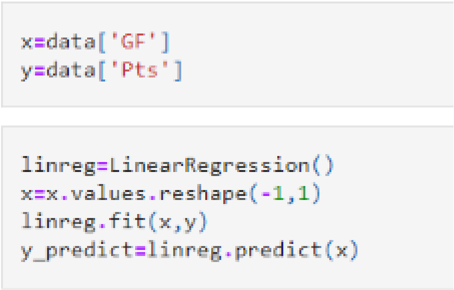
```
print("Slope=",slope,"intercept=",intercept,"R squared=",Rsqr)
```

```
Slope= [0.62036799] intercept= 52.268181818181816 R squared= 0.9344078657511434
```

**Should you buy attackers or defenders in the Premier League ?**

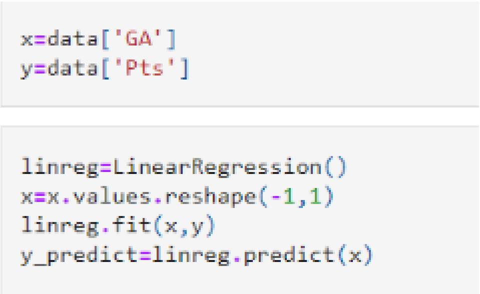
One of the most fiercely debated topics in the premier league is whether clubs should spend money on an attacking player or if the same money should be spent on buying a defender,given that both of them are equally capable in their respective roles. We tried to quantitatively analyse the problem with the help of linear regression.

The procedure is exactly the same but this time instead of the goal difference we replaced the X-axis variables with the goals scored and the goals conceded. In some cases, teams that finish the season with high points do not have more goals scored than those with fewer points, as they concede less goals. In some cases, the teams that finish the season high have more goals conceded, as they score more to compensate for the conceded goals.



```
x=data['GF']
y=data['Pts']
```

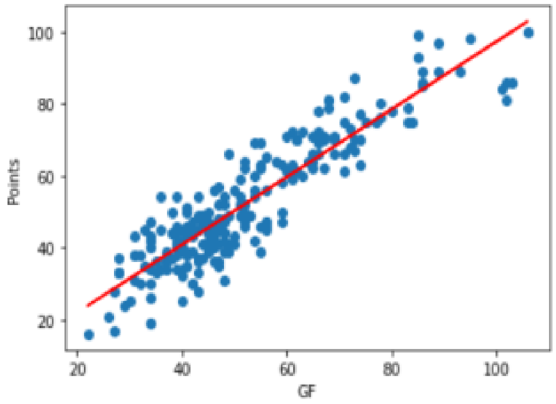
```
linreg=LinearRegression()
x=x.values.reshape(-1,1)
linreg.fit(x,y)
y_predict=linreg.predict(x)
```



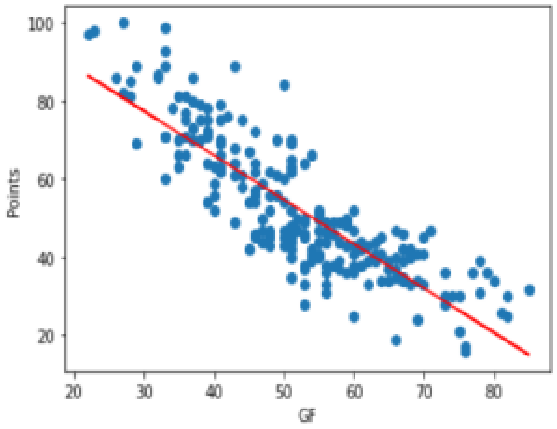
```
x=data['GA']
y=data['Pts']
```

```
linreg=LinearRegression()
x=x.values.reshape(-1,1)
linreg.fit(x,y)
y_predict=linreg.predict(x)
```

After importing each data point to the plot, we observed a linear relationship. But the correlation was weaker as compared to before. Both of the plots are shown below:

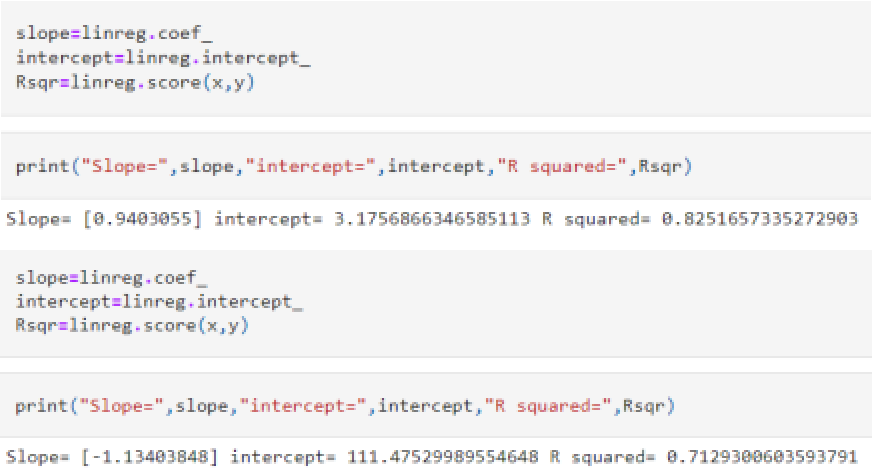


**A:Goals Scored Vs Points**



**B:Goals Conceded Vs Points**

Slope of the best fit line in A is positive as more goals scored should lead to higher points. Similarly, slope of best fit line in B is negative as less goals conceded should also lead to higher points. We observed α to be 0.94 and -1.14 in A and B respectively. From this, we can infer that one or more goals scored can increase the final points of a team by 0.94 whereas one more goal less conceded can increase the final points of the team by 1.14.



```
slope=linreg.coef_
intercept=linreg.intercept_
Rsqr=linreg.score(x,y)
```

```
print("Slope=",slope,"intercept=",intercept,"R squared=",Rsqr)
```

```
Slope= [0.9403055] intercept= 3.1756866346585113 R squared= 0.8251657335272903
```

```
slope=linreg.coef_
intercept=linreg.intercept_
Rsqr=linreg.score(x,y)
```

```
print("Slope=",slope,"intercept=",intercept,"R squared=",Rsqr)
```

```
Slope= [-1.13403848] intercept= 111.47529989554648 R squared= 0.7129300603593791
```

So, in terms of the result we get from applying linear regression to our data, a team is better off buying a defender than a scorer as less number of goals allowed is worth more points than the number of goals scored. This might not be applicable for other leagues. This is because many teams in the English Premier League have similar capabilities. Therefore, the goal difference is much smaller. As a result, better defense will allow teams to concede fewer goals, which will bring more points. Over the past few seasons, defenders are bringing more and more attacking returns as well. For many teams, defenders have more shots on target and chances created than attackers. So, it definitely makes sense to go for a defender rather than an attacker in the premier league.

However,this is not the case for other leagues in Europe.In other football leagues,like the Bundesliga or the Ligue 1, there is a huge gap in the level of dominant teams and the weaker teams. Just one team has been the champion for many years.In England, the top six teams have similar capabilities.As a result, the goal difference for each team will be small. A better defense will ensure the team allows fewer goals since their offense has been relatively strong, which brings more points considering the fact that a draw game is worth 1 point and a win game is worth 3 points. However, in the other leagues,the weaker teams allow plenty of goals and as a result, the goal difference is high. So, buying an attacking player is more efficient than buying a defender.