**Our goal is to stop yours!**

**Project description**

Input one of the 3 seasons with the suffix ‘.xlsx’ in the following format:

YYYY-YY.xlsx

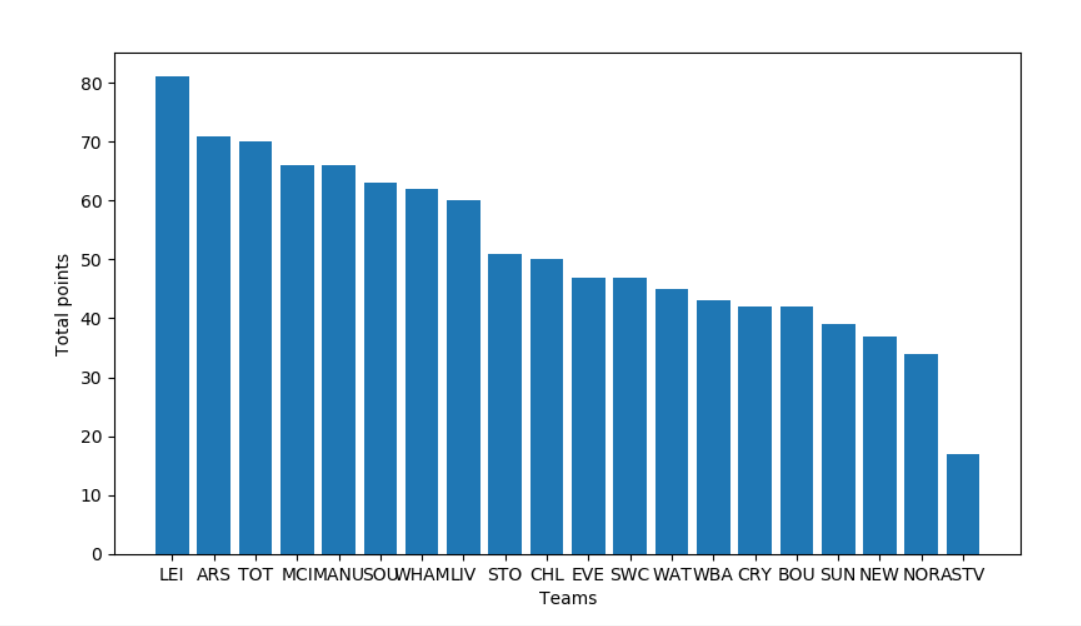
The 3 seasons available are: 2015-16, 2017-18, and 2017-18.

After taking the required season as input I have analyzed the data and printed a few stats from that season.

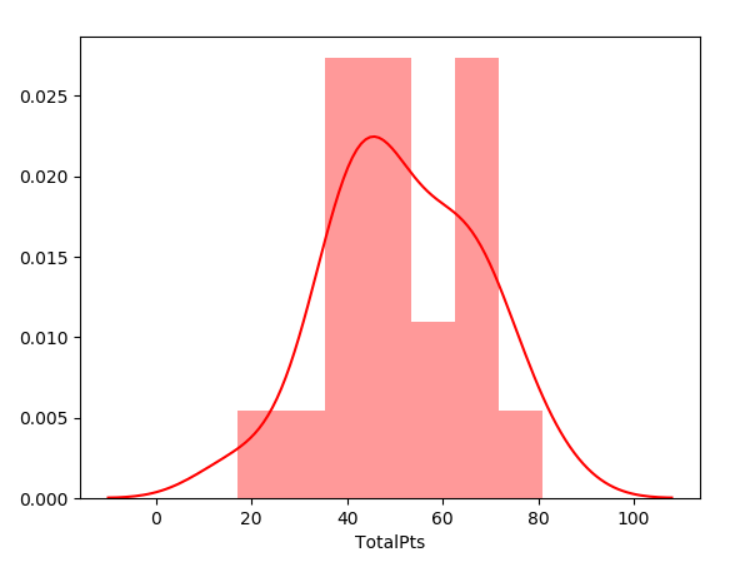
One thing I want to clarify before proceeding is that the term ‘variable’ I’ll be using throughout this project refers to the column names in the excel sheets such as Total points, Teams, Total wins etc and has no relation to the ‘variable’ term used in Python. Independent variables refer to those variables which are not dependent on other variables and the dependent variable is Total points for my project whose value depends on all the independent variables.

After this, I have plotted Bar graphs for the following observations:

* “Teams” vs “Total points”
* “Teams” vs “Total Wins”
* “Teams” vs “Total Draws”
* “Teams” vs “Total Losses”
* “Teams” vs “Total Goals Scored”
* “Teams” vs “Total Goals conceded”
* “Teams” vs “Total Fouls”
* “Teams” vs “Total Yellow cards”
* “Teams” vs “Total Red cards”
* “Teams” vs “Shot conversion ratio”



Then I have plotted a Histogram displaying the distribution of the “Total points”.



From the above graph you can see that this is a normal distribution. Hence, it can be used to fit a “Multiple regression model” and determine a relationship between “Total points” which is the dependent variable and the independent variables such as “Total Goals Conceded”,” Total Goals Scored”,” Fouls”, “Yellow cards” etc.(please refer to the annex at the end).

Before I move on further this is the basics of Multiple Linear Regression which helps you understand the rest of the program and interpret the results:

A line in a two dimensional or two-variable space is defined by the equation *Y=a+b\*X*; in full text: the Y variable can be expressed in terms of a constant (a) and a slope (b) times the X variable. The constant a is also referred to as the intercept, and the slope as the regression coefficient or B coefficient. For example, GPA may best be predicted as 1+.02\*IQ. Thus, knowing that a student has an IQ of 130 would lead us to predict that her GPA would be 3.6 (since, 1+.02\*130=3.6).

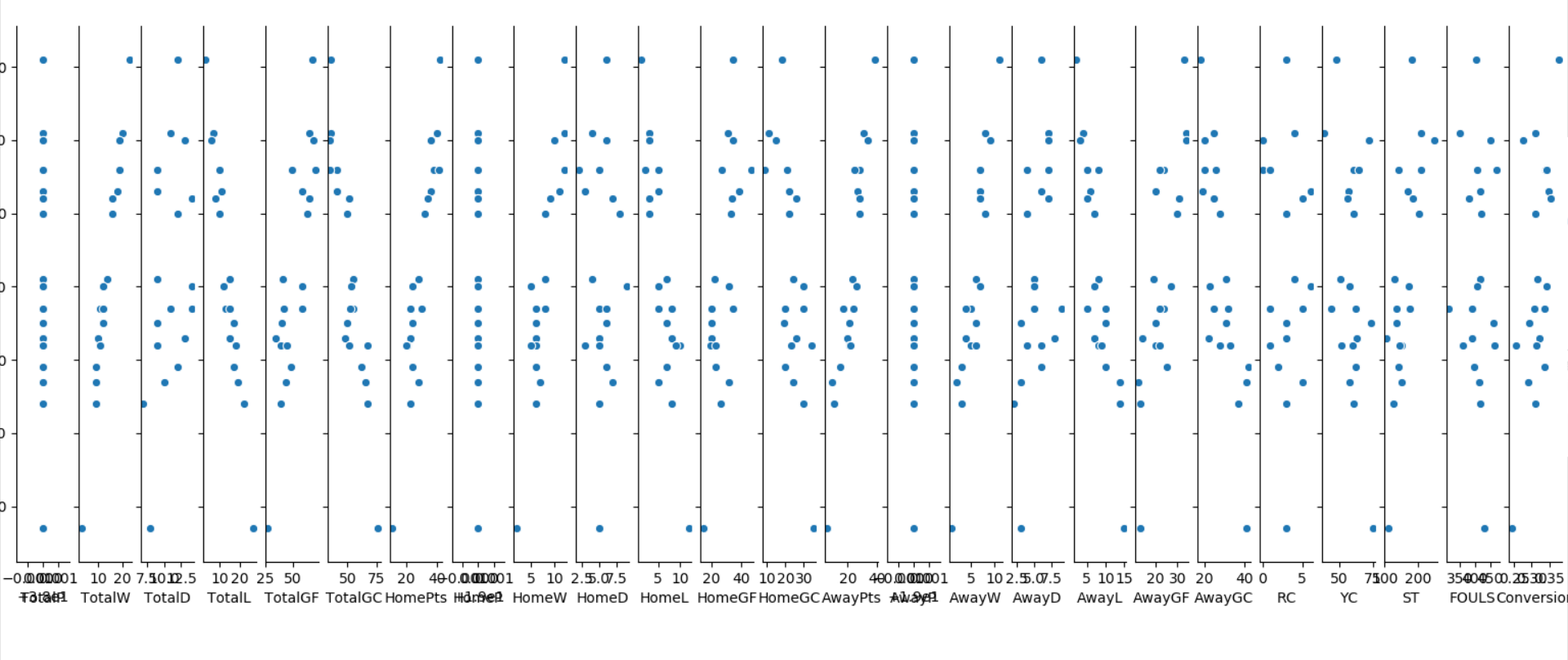
For example, the animation below shows a two dimensional regression equation plotted with three different confidence intervals (90%, 95% and 99%).

In a multiple linear regression, when there is more than one independent variable, the regression line cannot be visualized in the two dimensional space, but can be computed just as easily. For example, if in addition to IQ we had additional predictors of achievement (e.g., Motivation, Self- discipline) we could construct a linear equation containing all those variables. In general then, multiple regression procedures will estimate a linear equation of the form:

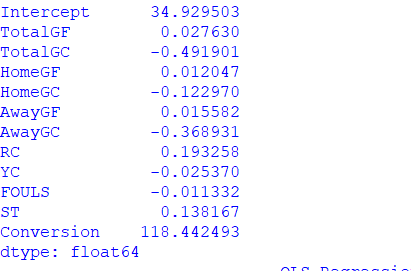
Y = a + b1\*X1 + b2\*X2 + ... + bp\*Xp

p : number of independent variables.

I plotted the 1 to 1 relationship between “Total points” and each of the independent variable used in the analysis.

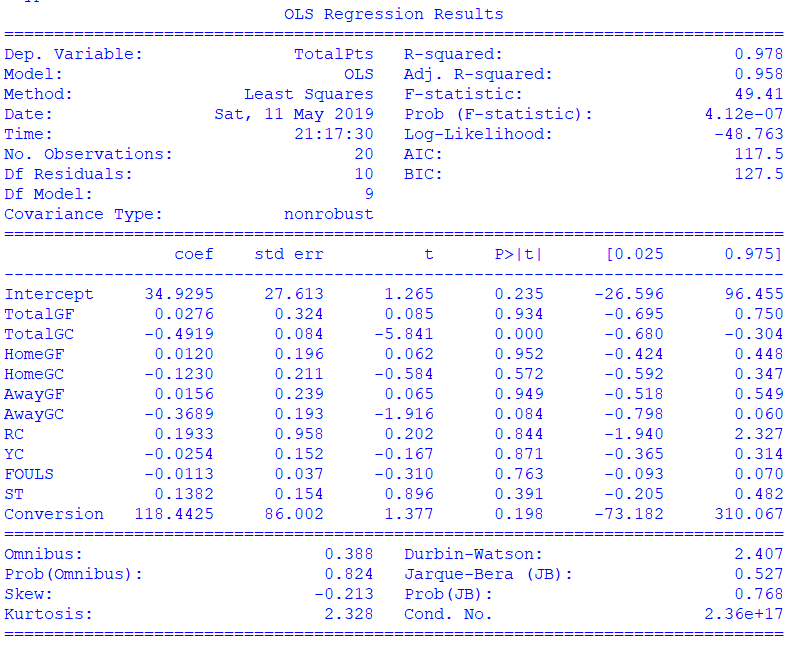


After this I derived the intercept of all the independent variables. PFA example:



As you can see from the picture above the intercept(a) is 34.9295, b0=0.027630, b1=-0.4919, b2=0.12047 and so on.

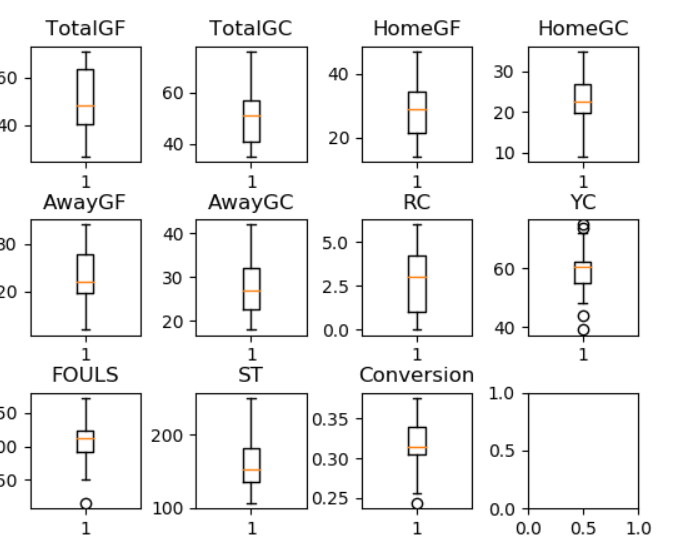
Next, I summarized the results:



There are a couple of inferences that can be made from the above figure:

* “Adj. R-squared” measures what percentage of the values of the Dependent variable (Total points in this case) can be explained by the model (Intercept and all the independent variables included). Hence, the more the percentage the better the predictions of the model. Initially when I run the model with all of the variables, the Adj R-squared value =1(100%) since, all the independent variables are included.
* P>|t| is the p-value of an independent variable. The p-value process determines a hypothesis and a hypothesis thesis is presented between each characteristic according to this hypothesis. The closer the p-value is to 0 the more significant the independent variable, and the closer its value to 1 the independent variable becomes insignificant. I’ve considered the cut-off p-value for this project as 0.1.
* The p-value of the variables such as “Total points”, “Total wins, draws, losses”, “Total Home wins, draws, losses” etc is displayed as 0.0000 which means that these variables are highly significant but it’s obvious that these variables are significant in determining the Total points. Hence, I have discarded these variables and considered variables such as “Total Goals Conceded”, “Yellow cards”, “Fouls” etc to predict the Total points.
* Discarding variables from the model will reduce the Adj. R-squared value. However, if the Adj. R-squared is still high and close to 1 then I can discard the variables and the model will still be able to predict the “Total points” accurately.
* Also, the F-stat value must be high (at least 10) for the independent variables in the final model to be significant. So, I have ensured that the final models have F-stat values of at least 22 or more.

I have also, plotted the box plots of the independent variables to identify the outliers. Box plot is used to find the lower quartile (25th percentile), median (50th percentile), upper quartile (75th percentile) and the maximum of a continuous variable.



After following the steps mentioned in the above points I have determined the final MLR models for each of the previous Premier league seasons (2015, 2016 and 2017) and have determine the final points each team will accumulate at the end of the season, hence predicting the winners and the relegation teams for this season.

**2015:**

The final model for this season is:

Total points = 130.3956 – (1.1317 \* Total goals conceded) – (0.38 \* Yellow cards) + (0.0042 \* Fouls)

Using the above equation, the Winner for the current season would be:

**Liverpool** with **92.75** points followed by **Manchester City** with **90.12** points.

The 3 teams to get relegated would be **Cardiff**, **Huddersfield** and **Fulham** with **30.34, 26.74** and **19.55** points respectively.

**2016:**

The final model for this season is:

Total points = 53.9391 – (0.4581 \* Total goals conceded) + (164.6356 \* Shot Conversion Ratio) – (0.9804 \* Away Goals conceded)

Using the above equation, the Winner for the current season would be:

**Liverpool** with **97.35** points followed by **Manchester City** with **94.19** points.

The 3 teams to get relegated would be **Cardiff, Fulham** and **Huddersfield** with **35.56, 13.07** and **6.95** points respectively.

**2017:**

The final model for this season is:

Total points = 110.2051 – (1.9985 \* Away goals conceded) Using the above equation, the Winner for the current season would be:

**Manchester City** with **90.22** points followed by **Liverpool** with **86.22** points.

The 3 teams to get relegated would be **Burnley**, **Huddersfield** and **Fulham** with **38.26, 22.27** and **20.27** points respectively.

The title race this season went down to the wire and Manchester City were crowned as champions on the final day (12th May).

As, you can see the Winners and the relegated teams are not the same in all 3 cases. I guess we must wait for the final day to find out who the champions will be.

I will be attaching the program, demo video and the excel sheets with the Final league table and the Season stats for the seasons 2015, 2016 and 2017. I will also be attaching the excel sheet with the stats for the current season.

The link to the video is: <https://www.youtube.com/watch?v=sE0FIoNf9Zg&feature=youtu.be>

**Annex:**