Author: Jaime Martin
Student No: 2914569

Subject: Big Data Analytics

Title: Asignment 3 – Part 1

Introduction:

To the end of the 2014/2015 season there have been 190006 FA-sanctioned league matches played in England across 14097 different matchdays. Here is the very first matchday on the 8th of September 1888.

	Date	Season	home	visitor	FT	hgoal	vgoal	division	tier	totgoal	goaldif	result
38	1888-09-08	1888	Bolton Wanderers	Derby County	3-6	3	6	1	1	9	-3	A
67	1888-09-08	1888	Everton	Accrington F.C.	2-1	2	1	1	1	3	1	H
93	1888-09-08	1888	Preston North End	Burnley	5-2	5	2	1	1	7	3	H
109	1888-09-08	1888	Stoke City	West Bromwich Albion	0-2	0	2	1	1	2	-2	A
123	1888-09-08	1888	Wolverhampton Wanderers	Aston Villa	1-1	1	1	1	1	2	0	D

The Variables:

In this data there are 12 variables and there are no missing values. Here is a short summary of the data in each column:

```
 $ Date    : Factor w/ 14097 levels "1888-09-08","1888-09-15",..: 17 23 32 15 6 20 24 7 36 14 ...
 $ Season  : int  1888 1888 1888 1888 1888 1888 1888 1888 1888 1888 ...
 $ home    : Factor w/ 142 levels "Aberdare Athletic",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ visitor : Factor w/ 142 levels "Aberdare Athletic",..: 10 15 17 26 48 51 95 102 122 132 ...
 $ FT      : Factor w/ 95 levels "0-0","0-1","0-10",..: 13 4 39 64 73 46 14 1 36 37 ...
 $ hgoal   : int  1 0 2 5 6 3 1 0 2 2 ...
 $ vgoal   : int  1 2 3 1 2 1 2 0 0 1 ...
 $ division: Factor w/ 6 levels "1","2","3","3a",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ tier    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ totgoal : int  2 2 5 6 8 4 3 0 2 3 ...
 $ goaldif : int  0 -2 -1 4 4 2 -1 0 2 1 ...
 $ result  : Factor w/ 3 levels "A","D","H": 2 1 1 3 3 3 1 2 3 3 ...
```

The columns that are Factors are categorical data. We must also convert Season and tier to factors as they should also be categorical.

For now the numeric variables are all discrete. These are hgoal, vgoal, totgoal and goaldif.

The next sections will explore the data in each column further by using the summary of the data frame.

Date:

Date is a factor representing the date that matches were played on. In part 2 this will be used to split seasons in half. The first half of the season is August to December and the second half is from January to June.

```
 1950-08-19:    46
 1950-08-26:    46
 1950-09-02:    46
 1950-09-09:    46
 1950-09-16:    46
 1950-09-23:    46
 (Other)   :189820
```

We can see that the dates with the most matches have 46 matches. This is the most possible as their are only 92 teams in the league at any one point. Also this started in 1950 when the leagues reach a total of 92 teams for the first time. Kickoffs on Saturday at 3pm used to be a tradition in English Football but since the introduction of Sky Sports in 1991/1992 games have been moved to various different days and times for television. Here are the last dates with 46 games played

Date	# Games
1991-05-11	46
1991-12-26	46
2003-12-26	46
2008-12-26	46
2013-12-26	46
2014-12-26	46

We have to go back to the last day of the 1990-91 Season to find a non Boxing day where all 92 league teams played.

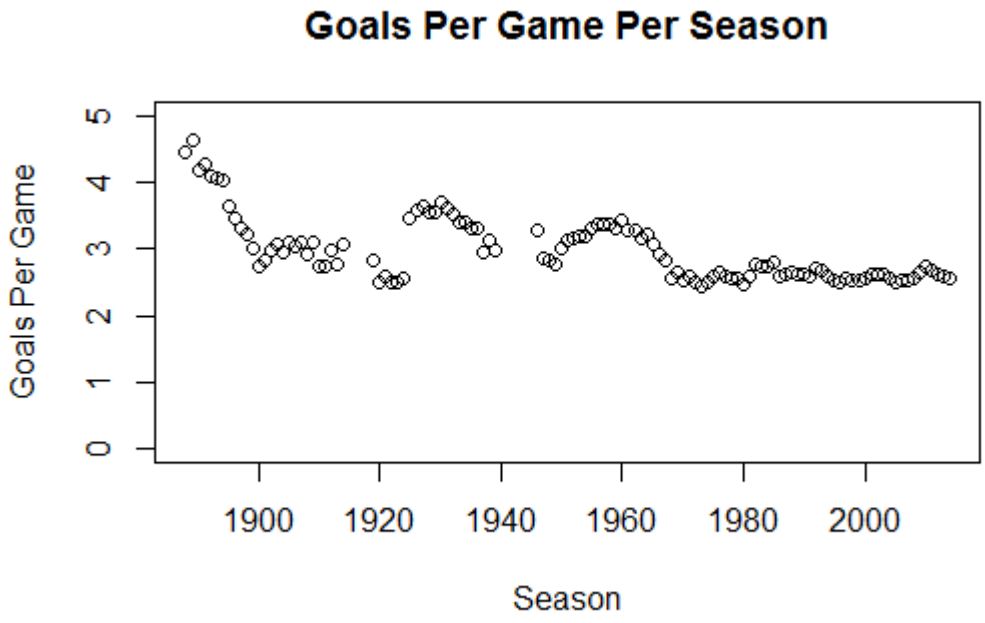
These are the 15 dates with the most goals scored (not normalized with number of games played):

Date	# Goals
1932-01-02	209
1936-02-01	209
1928-01-07	199
1930-11-01	196
1927-12-24	195
1928-02-04	195
1932-10-29	195
1925-09-26	194
1929-11-09	194
1931-11-07	190
1934-03-24	189
1927-11-05	188
1928-12-26	188
1933-01-07	188
1934-12-15	188

Interestingly despite not having matchdays with 46 games all of these occur between 1925 and 1936. Perhaps we need to look at scoring trends across seasons.

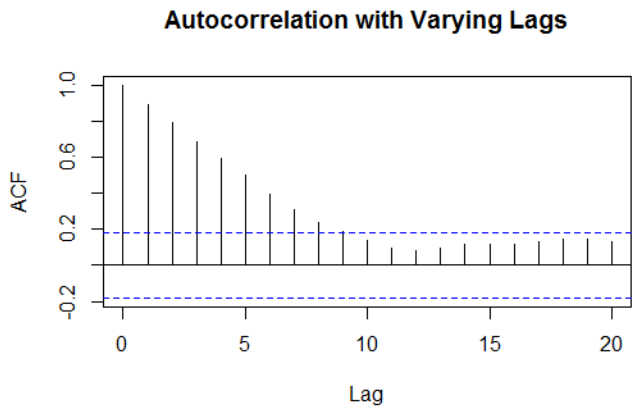
Season:

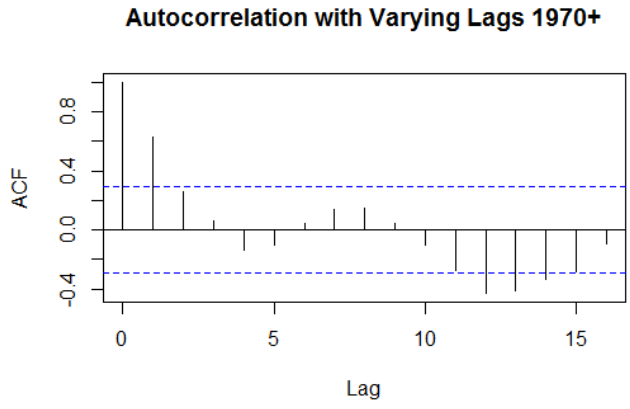
Seasons in soccer typically take place across two different years. The last season in this dataset is 2014/2015 and is represented by Season = 2014. The following graph shows the mean goals per game in all seasons.



At first it may appear that the line jumps about a bit before settling on a pretty straight trajectory. Note however the gaps just before 1920 and just after 1940. These gaps are where soccer was stopped for WWI and WWII. Young able-bodied men who would have been playing soccer would generally have been off fighting for King and country.

There are two humps after each gap. If we look at the seasons just before WWI it is fairly flat as well. So it seems that the rate of scoring does not change much. I checked this using autocorrelation where you check how predictive the rate of scoring in Season=t-1 is for Season=t. An interesting lesson was learned. I made two graphs one was the autocorrelation for all season and the second was the autocorrelation from 1970 on.





The correlations remain higher for longer when all seasons are taken into account. By splitting the data I was expecting the opposite as the values are closer together. But correlation is the ratio of explained variation. There is always some unexplained variation because of model underspecification. By making the variance very small, 0.0066 for 1970 on compared to 0.2238 for all seasons – 34 times bigger. So be careful using correlation when the slope is small – note the ratio between the slopes of linear models fitted to the data is 22.

Home & Visitor:

home			visitor		
Notts County	:	2402	Notts County	:	2402
Preston North End	:	2390	Preston North End	:	2389
Burnley	:	2373	Burnley	:	2373
Wolverhampton Wanderers	:	2359	Wolverhampton Wanderers	:	2360
Derby County	:	2337	Bury	:	2336
Bury	:	2335	Derby County	:	2336
(Other)	:	175900	(Other)	:	175900

Home and Visitor represent the teams playing at home and away respectively. The teams that have played the most games are those that have been around since the foundation of the leagues but have not always been in the the first tier where there are fewer teams and thus fewer games played. Note that Preston North End, Derby, Wolverhampton Wanderers and Bury have not played the same number of games. This may have been due to cancellations like weather or teams not fulfilling their schedule. This is something that may have to be taken into account when using small sample sizes – data is not unavailable in the sense that we use it is the no event occurred to gather data on.

FT:

This represent the score at full time. Their have been 95 different results in league history. Note that 1-0 is a home win and 0-1 is an away win and are not equivalent. Having noted the change in goals per season the sample was split into three buckets: pre 1920, 1920-1970, and after 1970. The top 10 results of all time were found and then for each of these results the observations in each bucket were found. The observations are expresedd as a percentage of expected observations (observed*100/expected).

Pre 1920:

1-1	1-0	2-1	2-0	0-0	0-1	1-2	2-2	3-1	3-0
79.6	95.7	94.3	106.8	88.0	87.4	89.2	83.4	111.0	121.5

Despite this being mostly a high scoring era there appears to be far fewer results where the away team scores (1-1, 2-1, 0-1, 1-2, 2-2) except for 3-1 but if we just look at the home team scoring 3 then 3-0 is relatively more frequent than 3-0. It may be worth looking at how home vs away goals have been distributed over time.

1920-1970:

1-1	1-0	2-1	2-0	0-0	0-1	1-2	2-2	3-1	3-0
92.8	87.5	95.7	96.1	84.8	80.9	92.2	102.2	113.4	105.7

There is no weird split in this era. There were more goals scored on average and thus the low-scoring games are less frequent and the higher scoring ones are more frequent.

After 1970:

1-1	1-0	2-1	2-0	0-0	0-1	1-2	2-2	3-1	3-0
110.6	112.5	105.2	102.3	116.5	120.2	109.4	101.2	85.4	90.5

This is the converse of the previous era: fewer goals are scored on average so lower-scoring result occur more often.

Hgoal, Vgoal, Totgoal, and Goaldif:

These variables represent goals scored by the home team, the away team, both teams, and home-away respectively. Following are the summary statistic for each of those variables for the three buckets discussed in the previous section. Note that the mean is always greater than the median indicating that the data is positively skewed. Which should be obvious as you cannot have negative goals.

Pre 1920:

hgoal	vgoal	totgoal	goaldif
Min. : 0.000	Min. : 0.000	Min. : 0.00	Min. :-10.0000
1st Qu.: 1.000	1st Qu.: 0.000	1st Qu.: 2.00	1st Qu.: 0.0000
Median : 2.000	Median : 1.000	Median : 3.00	Median : 1.0000
Mean : 2.006	Mean : 1.084	Mean : 3.09	Mean : 0.9226
3rd Qu.: 3.000	3rd Qu.: 2.000	3rd Qu.: 4.00	3rd Qu.: 2.0000
Max. :12.000	Max. :10.000	Max. :14.00	Max. : 12.0000

1920-1970:

hgoal	vgoal	totgoal	goaldif
Min. : 0.000	Min. :0.000	Min. : 0.000	Min. :-9.00
1st Qu.: 1.000	1st Qu.:0.000	1st Qu.: 2.000	1st Qu.: 0.00
Median : 2.000	Median :1.000	Median : 3.000	Median : 1.00
Mean : 1.978	Mean :1.168	Mean : 3.146	Mean : 0.81
3rd Qu.: 3.000	3rd Qu.:2.000	3rd Qu.: 4.000	3rd Qu.: 2.00
Max. :13.000	Max. :9.000	Max. :17.000	Max. :13.00

After 1970:

hgoal	vgoal	totgoal	goaldif
Min. : 0.000	Min. :0.000	Min. : 0.000	Min. :-8.0000
1st Qu.: 1.000	1st Qu.:0.000	1st Qu.: 1.000	1st Qu.:-1.0000
Median : 1.000	Median :1.000	Median : 2.000	Median : 0.0000
Mean : 1.531	Mean :1.067	Mean : 2.598	Mean : 0.4635
3rd Qu.: 2.000	3rd Qu.:2.000	3rd Qu.: 4.000	3rd Qu.: 1.0000
Max. :10.000	Max. :9.000	Max. :12.000	Max. :10.0000

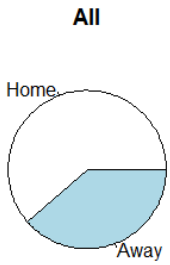
We wondered in the last section how home goals were distributed relative to away goals. On the next page are pie charts based on the percentage of total goals that were scored by the home team and the away team. You can see that the slice for the home team is greatest pre 1920 but I did not find it easy to spot until I knew what I was looking for.

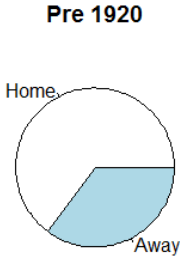
We can see it more clearly by looking at the mean goal difference. This would not bue useful if the mean of home goals or away goals varied wildly i.e. pre 1920 averaged 10 goals a game and after 1970 averaged 2 goals a game. We know this is not the case so we get the following results

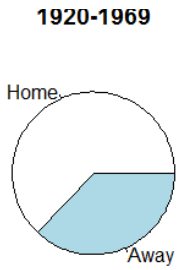
Pre 1920: 0.92
1920-1970: 0.81
After 1970: 0.46

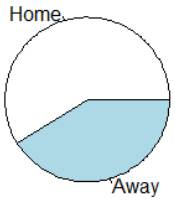
The mean difference has halved but the percentages

Pre 1920: 0.65 v 0.35
1920-1970: 0.63 v 0.37
After 1970: 0.59 v 0.41



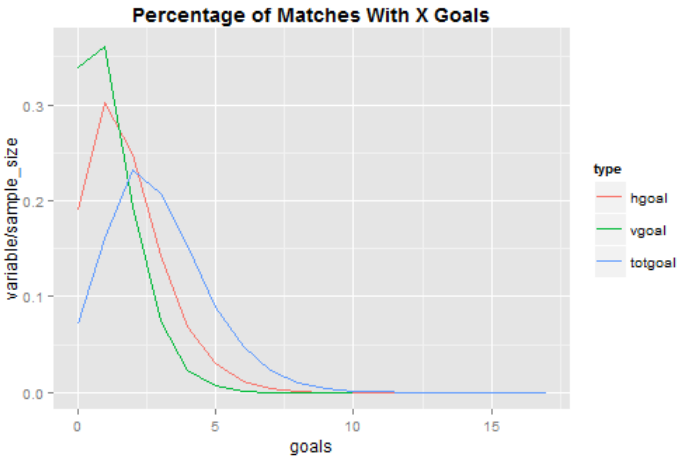


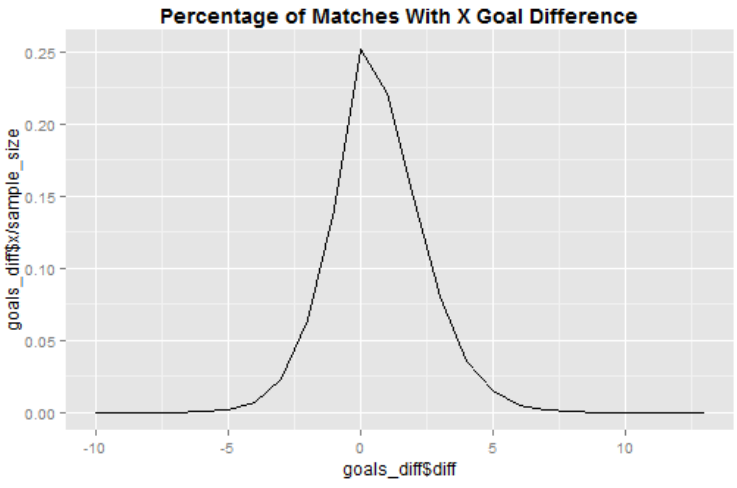




Here is a plot of the distribution of goals scored home,away, and by both teams:

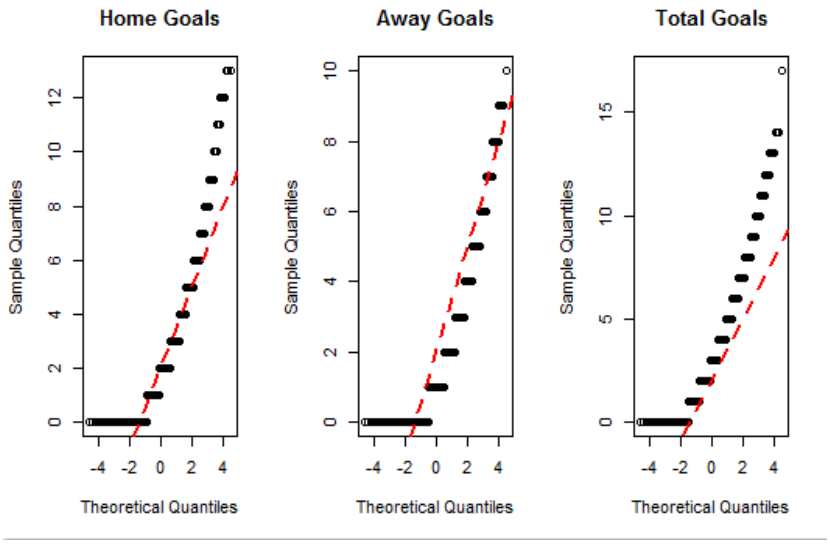
The distribution of totgoal certainly seems to be normal. That of hgoal and vgoal is preventented from being symmetric as a normal distribution requires. Below we have the distribution of goaldif. It is indeed symmetric and looks like the normal distribution. Since the difference between two normal distributions is itself normal this offers some evidence that hgoal and vgoal are at least approximately normal.





In the quantile-quantile plot we can see there is a long black line at zero this is where negative values should appear which is not a possible outcome in a soccer match.

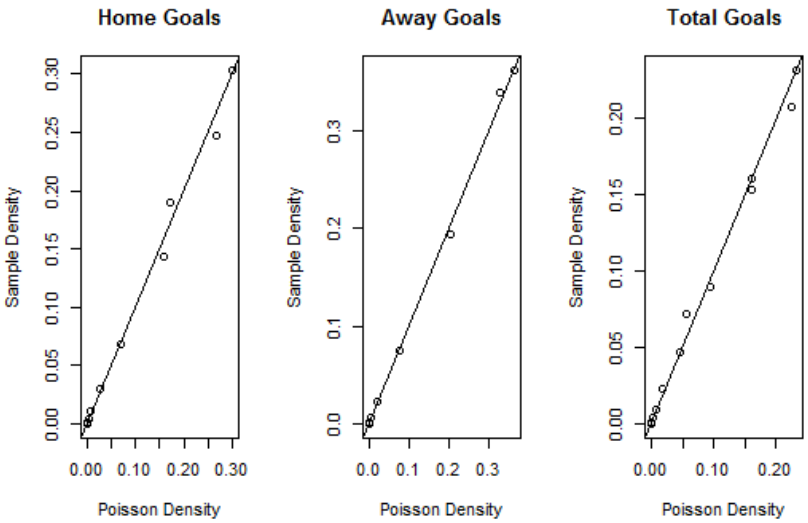
It would appear that away goals fits the best. With a normal distribution failing to predict very high goals scored by home teams. This may be because there is one event that can thoroughly change the rate of goals scored in a soccer match. That is a red card. It is possible that these big scores involved multiple red card for the away team with referees, feeling pressure from the home crowd, showing a bias towards the home team.



Here we plot the quantiles against the Poisson Distribution with all the points very close to the 45 degree line. Being a Poisson distribution is very helpful for further analysis.

First of all the mean of the distribution is equal to the variance so we do not have to find two parameters as we do in the normal distribution.

Secondly Poisson distributions can be linearly combined so we do not have to worry about mixing distributions.



Tier and Division:

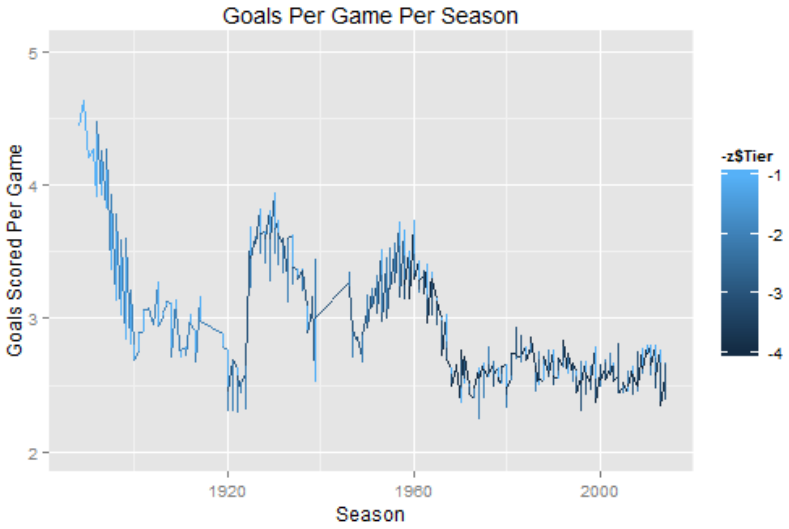
Tier represents the ordinal rank of the different leagues in the Football League. At present we have the following

League Name	Tier
Premier League	1
Championship	2
League 1	3
League 2	4

Division represents the fact that there used to be North and South Divisions that were of equals status (represented by 3a and 3b in the dataset). This may prove superfluous and could be removed from the dataset in part 2.

The plot on the right shows the Goals Per Game Per Season metric that was used in the Season section above. This time it is split into four lines that are coloured by tier.

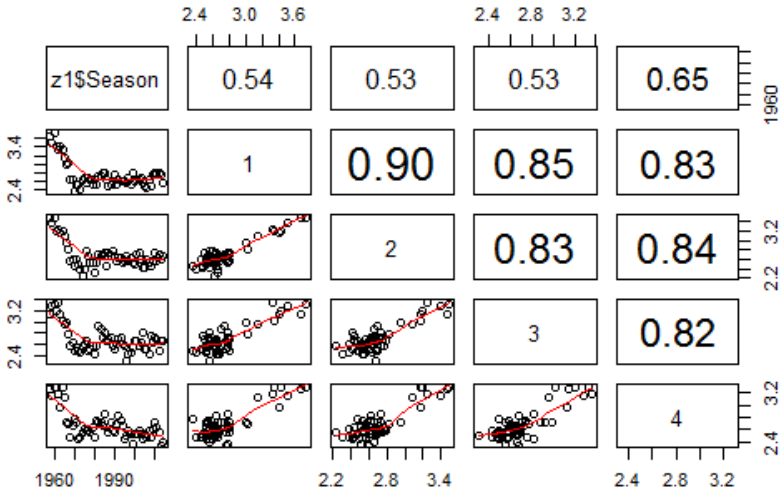
It can be seen that all four divisions generally have scoring rates that are close to each other as it is hard to discern the lines on the graph. This seems like an analysis more suited to an ANOVA test so we will leave it to part 2.



Here we have a matrix of scatterplots and correlations for Seasons and each of the tiers. The data only uses Seasons from 1958 so that we can calculate the correlations.

The first column contains the correlations for Seasons and each tier and they clearly look similar. In the other columns in the lower panel we see goals in each tier correlated with each other and this produces a smoothed line that is roughly along the diagonal.

Finally it can be seen that correlations between goals in each tier are very high.



Result:

Result is a factor with 3 levels: 'H'. 'A', and 'D'. These represent the three states a match can finish in. This may enable easy calculation of league tables for home and away form. To create the full league tables

New Data:

Now that we understand the data that is already in the data set we can start to build more interesting statistics. First we must note that when we add season data to be fair we should translate statistics as if 38 games were played. So, for example, if a team had a +50 goal difference in 46 games then we divide that by 46 and multiply by 38.

Here then are the best goal differences per 38 games in league history (min. 10 games played):

home	Season	tier	gd
Lincoln City	1975	4	72
Chelsea	2009	1	71
Liverpool	1978	1	69
Reading	2005	2	67
Notts County	2009	4	65
Manchester City	2013	1	65
Manchester City	2011	1	64
Sunderland	1998	2	63
Liverpool	1987	1	63
Bristol City	2014	3	58
Wolverhampton Wan...	2013	3	58
Manchester United	2009	1	58
Fulham	2000	2	58
Manchester United	2007	1	58
York City	1983	4	57
Chelsea	2004	1	57
Milton Keynes Dons	2014	3	57
Manchester United	2011	1	56
Arsenal	1990	1	56
Manchester United	2006	1	56

And the worst (min. 10 games played):

home	Season	tier	gd
Doncaster Rovers	1997	4	-83
Newport County	1987	4	-70
Derby County	2007	1	-69
Stoke City	1984	1	-67
Workington	1976	4	-61
Stockport County	2001	2	-60
Stockport County	2009	3	-60
Huddersfield Town	1987	2	-59
Cambridge United	1984	3	-58
Ipswich Town	1994	1	-57
Workington	1975	4	-57
Rochdale	1973	3	-56
Crewe Alexandra	1981	4	-55
Blackpool	2014	2	-55
Torquay United	1995	4	-54
Darlington	2009	4	-54
West Bromwich Albion	1985	1	-54
Swindon Town	1993	1	-53
Chester	1992	3	-53
Wolverhampton Wan...	1983	1	-53

Of course what every one cares about in football is points as they determine championships. Here are the best teams by Ppg (points per 38 games).

home	Season	tier	GP	W	D	L	GF	GA	GD	GFpg	GApg	GDpg	Ppg	GDRatio
Chelsea	2004	1	38	29.0	8.0	1.0	72	15	57	72.0	15.0	57.0	95.0	4.8
Chelsea	2005	1	38	29.0	4.0	5.0	72	22	50	72.0	22.0	50.0	91.0	3.27
Manchester United	1999	1	38	28.0	7.0	3.0	97	45	52	97.0	45.0	52.0	91.0	2.16
Manchester United	2008	1	38	28.0	6.0	4.0	68	24	44	68.0	24.0	44.0	90.0	2.83
Arsenal	2003	1	38	26.0	12.0	0.0	73	26	47	73.0	26.0	47.0	90.0	2.81
Manchester United	2006	1	38	28.0	5.0	5.0	83	27	56	83.0	27.0	56.0	89.0	3.07
Manchester United	2012	1	38	28.0	5.0	5.0	86	43	43	86.0	43.0	43.0	89.0	2.0
Manchester United	2011	1	38	28.0	5.0	5.0	89	33	56	89.0	33.0	56.0	89.0	2.7
Manchester City	2011	1	38	28.0	5.0	5.0	93	29	64	93.0	29.0	64.0	89.0	3.21
Liverpool	1978	1	42	30.0	8.0	4.0	85	16	69	77.0	14.0	62.0	89.0	5.5
Reading	2005	2	46	31.0	13.0	2.0	99	32	67	82.0	26.0	55.0	88.0	3.15
Lincoln City	1975	4	46	32.0	10.0	4.0	111	39	72	92.0	32.0	59.0	88.0	2.88
Arsenal	2001	1	38	26.0	9.0	3.0	79	36	43	79.0	36.0	43.0	87.0	2.19
Sunderland	1998	2	46	31.0	12.0	3.0	91	28	63	75.0	23.0	52.0	87.0	3.26
Manchester United	2007	1	38	27.0	6.0	5.0	80	22	58	80.0	22.0	58.0	87.0	3.64
Chelsea	2014	1	38	26.0	9.0	3.0	73	32	41	73.0	32.0	41.0	87.0	2.28
Manchester City	2013	1	38	27.0	5.0	6.0	102	37	65	102.0	37.0	65.0	86.0	2.76
Liverpool	1987	1	40	26.0	12.0	2.0	87	24	63	83.0	23.0	60.0	86.0	3.61
Liverpool	2008	1	38	25.0	11.0	2.0	77	27	50	77.0	27.0	50.0	86.0	2.85
Chelsea	2009	1	38	27.0	5.0	6.0	103	32	71	103.0	32.0	71.0	86.0	3.22

This will be where we search for correlations in part 2. The interesting explanatory variables are GFpg (Goals for per 38 games), GApg (Goals against per 38 games), GDpg (Goal Difference per 38 games), and GDRatio (Goals For/ Goals Against)

Furthermore correlations are only interesting when they are predictive so we will divide the seaso into first and second halves. Obviously at the end of the season points correlates perfectly with points. But it is often said that the best teams are the ones with the best goal difference. So which is more predictive of final points? Is it points at halfway or goal difference at halfway.

Finally we show the table from the 2014/15 season to show check that the code is working.

home	Season	tier	GP	W	D	L	GF	GA	GD	GFpg	GApg	GDpg	Ppg	GDRatio
Chelsea	2014	1	38	26.0	9.0	3.0	73	32	41	73.0	32.0	41.0	87.0	2.28
Manchester City	2014	1	38	24.0	7.0	7.0	83	38	45	83.0	38.0	45.0	79.0	2.18
Arsenal	2014	1	38	22.0	9.0	7.0	71	36	35	71.0	36.0	35.0	75.0	1.97
Manchester United	2014	1	38	20.0	10.0	8.0	62	37	25	62.0	37.0	25.0	70.0	1.68
Tottenham Hotspur	2014	1	38	19.0	7.0	12.0	58	53	5	58.0	53.0	5.0	64.0	1.09
Liverpool	2014	1	38	18.0	8.0	12.0	52	48	4	52.0	48.0	4.0	62.0	1.08
Southampton	2014	1	38	18.0	6.0	14.0	54	33	21	54.0	33.0	21.0	60.0	1.64
Swansea City	2014	1	38	16.0	8.0	14.0	46	49	-3	46.0	49.0	-3.0	56.0	0.94
Stoke City	2014	1	38	15.0	9.0	14.0	48	45	3	48.0	45.0	3.0	54.0	1.07
Crystal Palace	2014	1	38	13.0	9.0	16.0	47	51	-4	47.0	51.0	-4.0	48.0	0.92
West Ham United	2014	1	38	12.0	11.0	15.0	44	47	-3	44.0	47.0	-3.0	47.0	0.94
Everton	2014	1	38	12.0	11.0	15.0	48	50	-2	48.0	50.0	-2.0	47.0	0.96
West Bromwich Albion	2014	1	38	11.0	11.0	16.0	38	51	-13	38.0	51.0	-13.0	44.0	0.75
Leicester City	2014	1	38	11.0	8.0	19.0	46	55	-9	46.0	55.0	-9.0	41.0	0.84
Newcastle United	2014	1	38	10.0	9.0	19.0	40	63	-23	40.0	63.0	-23.0	39.0	0.63
Aston Villa	2014	1	38	10.0	8.0	20.0	31	57	-26	31.0	57.0	-26.0	38.0	0.54
Sunderland	2014	1	38	7.0	17.0	14.0	31	53	-22	31.0	53.0	-22.0	38.0	0.58
Hull City	2014	1	38	8.0	11.0	19.0	33	51	-18	33.0	51.0	-18.0	35.0	0.65
Burnley	2014	1	38	7.0	12.0	19.0	28	53	-25	28.0	53.0	-25.0	33.0	0.53
Queens Park Rangers	2014	1	38	8.0	6.0	24.0	42	73	-31	42.0	73.0	-31.0	30.0	0.58