



Premier League Analytics

BT1101 Project

Group 3

Ang Zhen Xuan A0139569L
Selina Wong Jinen A0142657B
Tan Qing Yang A0139508Y
Tham Shi Yuan A0093999Y
Yu Zongdong A0140018X

BT1101 Project: Premier League Analytics

Main Objective

To assist the club manager/coach to identify the strengths of the current teams in terms of the total number of passes, crosses and shots in each match, as well as to identify the better players within each team. We have two broad aims. One of which is to identify the top attributes which contributed to match winning across all teams so that the team managers may consider focusing on these particular techniques in their trainings. We also examined the various attributes which would attract more fans to watch their games so as to aid the club manager in generating greater revenue.

Objective of Descriptive Analytics

To determine which are the best players and best attributes which contributed to the top teams winning. We did this through different types of analysis. Firstly, we provided an overall summary of the team variables in the dataset which would be significant in our project. Secondly, we identified the players who have a higher win rate among matches that they played at least 75 minutes for. Next, we identified the best 4 statistics of the Top 2 Teams and Top 4 Teams in terms of rank, to examine the association between these attributes and the rank of the Teams in Premier League. Finally, we also looked at a player level. In particular, the Strikers of the Top 2 Teams in the Premier League to further identify the best qualities possessed by Strikers.

Importance

To identify the best qualities in the Top teams while taking into account the Striker's capabilities. This helps team managers identify their team's weaknesses compared to that of the Top Teams, to ensure that they are able to make necessary preparations before the matches. Also, by finding out which are the better players, the football manager would be better informed in his decision-making of the selection of players for crucial football matches in order to increase the chances of winning the game.

Objective of Predictive Analytics

Firstly, using linear regression, we aim to determine the attributes which make a football match interesting to attract more fans to watch, using match data from “tiki-taka” period of popularity. Secondly, in data-mining, we used logistic regression to identify factors that improve the chances of winning and examine these traits in the team.

Importance

The coach could determine the philosophy the team plays with, as it affects the amount of fans attracted and their revenue earned. With the predictive model we establish from running the regression, the team can focus on those attributes to establish the Tiki-Taka playing style and at the same time, tailor their training to achieve an attractive style of playing.

Data Source

Most of the data were obtained from OPTA for the purpose of performance analytics research as part of the MCFC Analytics project. The package contains guides such as event definitions which explain professional terms used in the soccer which assisted us in better understanding the data. Besides, the Premier league handbook for the season 2011-2012, MCFC Analysis Opta Formation Codes and Match by Match analysis data excel sheet were also utilized in our project. We also retrieved the 2011 – 2012 League table from the official website of the Barclays Premier League (ESPN FC, n.d.).

Tables

2011 / 2012 Season

Barclays Premier League

OVERALL

LIVE

		OVERALL						HOME						AWAY							
POS	TEAM	P	W	D	L	F	A	W	D	L	F	A	W	D	L	F	A	GD	PTS		
▶ 1	Manchester City	38	28	5	5	93	29	18	1	0	55	12	10	4	5	38	17	64	89		
▶ 2	Manchester United	38	28	5	5	89	33	15	2	2	52	19	13	3	3	37	14	56	89		
▶ 3	Arsenal	38	21	7	10	74	49	12	4	3	39	17	9	3	7	35	32	25	70		
▶ 4	Tottenham Hotspur	38	20	9	9	66	41	13	3	3	39	17	7	6	6	27	24	25	69		
▶ 5	Newcastle United	38	19	8	11	56	51	11	5	3	29	17	8	3	8	27	34	5	65		
▶ 6	Chelsea	38	18	10	10	65	46	12	3	4	41	24	6	7	6	24	22	19	64		
▶ 7	Everton	38	15	11	12	50	40	10	3	6	28	15	5	8	6	22	25	10	56		
▶ 8	Liverpool	38	14	10	14	47	40	6	9	4	24	16	8	1	10	23	24	7	52		
▶ 9	Fulham	38	14	10	14	48	51	10	5	4	36	26	4	5	10	12	25	-3	52		

Figure 1: League Table for 2011-2012

Suitability of Datasets

The datasets was used to differentiate the top, middle and bottom teams, which was required in our descriptive analysis and predictive analysis. Data for each team and every match of the season, which will culminate in a fair and easy way to analyse a team compared to the other teams in the same season 2011-12.

Biasness in Dataset

Media bias may occur even though factual stats were collected from every match played through the official data. The media is a powerful tool, influence and reach worldwide is strong, and may influence people's perception of a certain player, affecting the value of said player negatively or positively.

Descriptive Analytics

Summary Statistics

Row.Labels		Sum.of.Total.Successful.Passes.All		
Arsenal	: 1	Min.	:	7993
Aston Villa	: 1	1st Qu.:	:	10816
Blackburn Rovers	: 1	Median	:	12249
Bolton Wanderers	: 1	Mean	:	25647
Chelsea	: 1	3rd Qu.:	:	17217
Everton	: 1	Max.	:	269297
(Other)	:15			
Sum.of.Successful.Short.Passes		Sum.of.Successful.Long.Passes		
Min.	: 6337	Min.	:	828
1st Qu.:	9369	1st Qu.:	:	1023
Median	: 10567	Median	:	1097
Mean	: 22434	Mean	:	2192
3rd Qu.:	15039	3rd Qu.:	:	1293
Max.	:235560	Max.	:	23014
Sum.of.Successful.Ball.Touch		Sum.of.Saves.Made		
Sum.of.Goals.Conceded				
Min.	: 293.0	Min.	:	84.0
1st Qu.:	339.0	1st Qu.:	:	106.0
Median	: 393.0	Median	:	119.0
Mean	: 761.4	Mean	:	226.5
3rd Qu.:	441.0	3rd Qu.:	:	136.0
Max.	:7995.0	Max.	:	2378.0
Sum.of.Interceptions		Sum.of.Tackles.Won		
Sum.of.Successful.Dribbles				
Min.	: 461	Min.	:	463
1st Qu.:	574	1st Qu.:	:	513
Median	: 617	Median	:	535
Mean	: 1156	Mean	:	1016
3rd Qu.:	652	3rd Qu.:	:	553
Max.	:12139	Max.	:	10663
		Min.	:	114.0
		1st Qu.:	:	214.0
		Median	:	239.0
		Mean	:	470.2
		3rd Qu.:	:	305.0
		Max.	:	4937.0

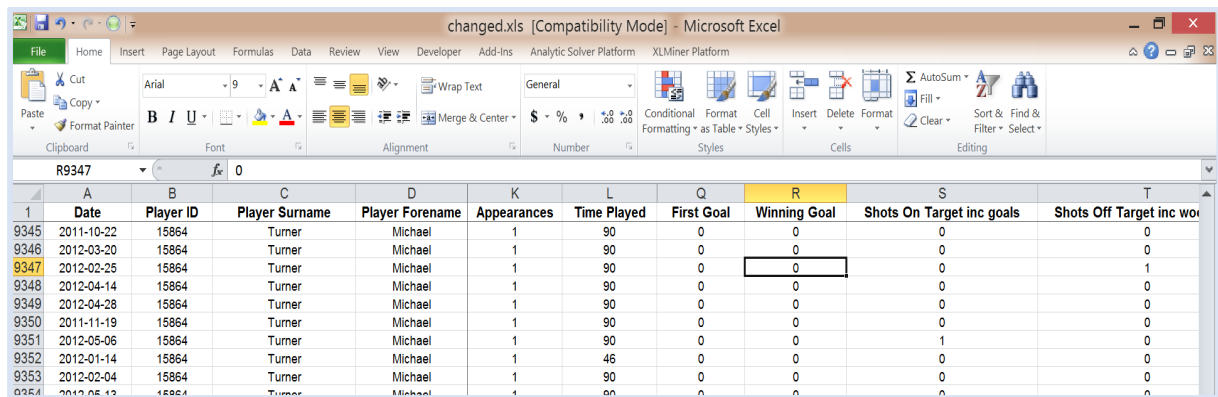
Using the Summary() function in R, we obtained the summary statistics for Sum of Total Successful Passes All, Sum of Successful Short Passes, Sum of Successful Long Passes, Sum of Successful Ball Touch, Sum of Saves Made, Sum of Goals Conceded, Sum of Interceptions, Sum of Tackles won and Sum of Successful Dribbles for all 20 teams in the Premier League in 2011-12.

Descriptive Analytics

Contributions of Non-Key Players

For clubs to gain a better understanding of their current team, we focused our analysis on non-key players that only played between 10 – 19 matches which they played at least 75 minutes in. We will then calculate the win rate for those players to determine if the player were performing well for the team.

In our analysis, we recognize that the teammates and the opponents that these players played with or against are different in each match. Hence, if a particular player has a higher winning rate, we use this result as an indicator to further look into the player in the future.



	A	B	C	D	K	L	Q	R	S	T
1	Date	Player ID	Player Surname	Player Forename	Appearances	Time Played	First Goal	Winning Goal	Shots On Target inc goals	Shots Off Target inc goals
9345	2011-10-22	15864	Turner	Michael	1	90	0	0	0	0
9346	2012-03-20	15864	Turner	Michael	1	90	0	0	0	0
9347	2012-02-25	15864	Turner	Michael	1	90	0	0	0	1
9348	2012-04-14	15864	Turner	Michael	1	90	0	0	0	0
9349	2012-04-28	15864	Turner	Michael	1	90	0	0	0	0
9350	2011-11-19	15864	Turner	Michael	1	90	0	0	0	0
9351	2012-05-06	15864	Turner	Michael	1	90	0	0	1	0
9352	2012-01-14	15864	Turner	Michael	1	46	0	0	0	0
9353	2012-02-04	15864	Turner	Michael	1	90	0	0	0	0
9354	2012-05-12	15864	Turner	Michael	1	90	0	0	0	0

Figure 2: Excel Sheet Data on Players

The data provided to us did not have the required information needed to carry out the descriptive analysis, such as Played Time over 75 minutes, Over 75 & won and the matches that the non-key players had played and won. We had to extract and consolidate the data in order to perform the analysis.

```
def read(filename):
    with open (filename, 'r') as f:
        all_ = ()
        for line in f:
            line = line[:-1]
            Date, PlayerID, TeamId, WinningGoal = line.split(',')
            data_point = (Date, PlayerID, TeamId, WinningGoal)
            all_ += (data_point,)
        return all_

data = read('Win Match Player.txt')
#data = read('test.txt')
data = data[1:]

win_match = ()

for data_point in data:
    for data_point2 in data:
        if data_point[0] == data_point2[0] and data_point[2] == data_point2[2] and data_point[3] == '1':
            win_match += (1,)
            break
    else:
        win_match += (0,)

print(len(win_match))
```

Figure 3: Python code for won matches

To attain the aforementioned information, we filtered date, playerId, TeamID and winningGoal into a separate excel file, then converted it into a csv file and finally imported it onto python.

Because the dataset only consists of the winning goal made by one player on the team as the indicator of whether a match was won or not, we had to iterate through the entire list of players to identify if the player played in a match he won. After obtaining this information, we created a dummy variable in the original excel file to indicate this as can be seen by the "Match Won" column in the image below.

	A	B	C	K	L	M	N	U	
	Date	Player ID	Player Surname	Appearance	Time Played	PlayedTime over 75 mins	Over 75 & Won	Match Won	Shots On T
3	2011-08-13	27450	Agbonlahor	1	90	1	0		0
4	2011-08-13	21094	Agger	1	90	1	1		1
5	2011-08-13	28491	Al-Habsi	1	90	1	0		0
9	2011-08-13	17997	Bardsley	1	90	1	0		0
11	2011-08-13	15276	Barton	1	90	1	1		1
12	2011-08-13	10738	Bent	1	90	1	0		0
13	2011-08-13	19008	Berra	1	90	1	0		0
17	2011-08-13	4255	Bothroyd	1	90	1	1		1
18	2011-08-13	3296	Boyce	1	90	1	0		0

Figure 4: Excel Sheet Data on Players who played over 75 minutes

Our next step was filtering out players who had played at least 75 minutes, using the "If" function on Excel. If the time played is more than or equal to 75 then the 'played time over 75 minutes' would be assigned 1, otherwise, it would be assigned 0. We created another column 'Over 75 minutes and Won' to determine games in which players had played more than 75 minutes and won. If the sum of 'Playedtime over 75 minutes' and 'Match Won' equals to 2 then the new column, 'Over 75 minutes and Won' would be assigned 1.

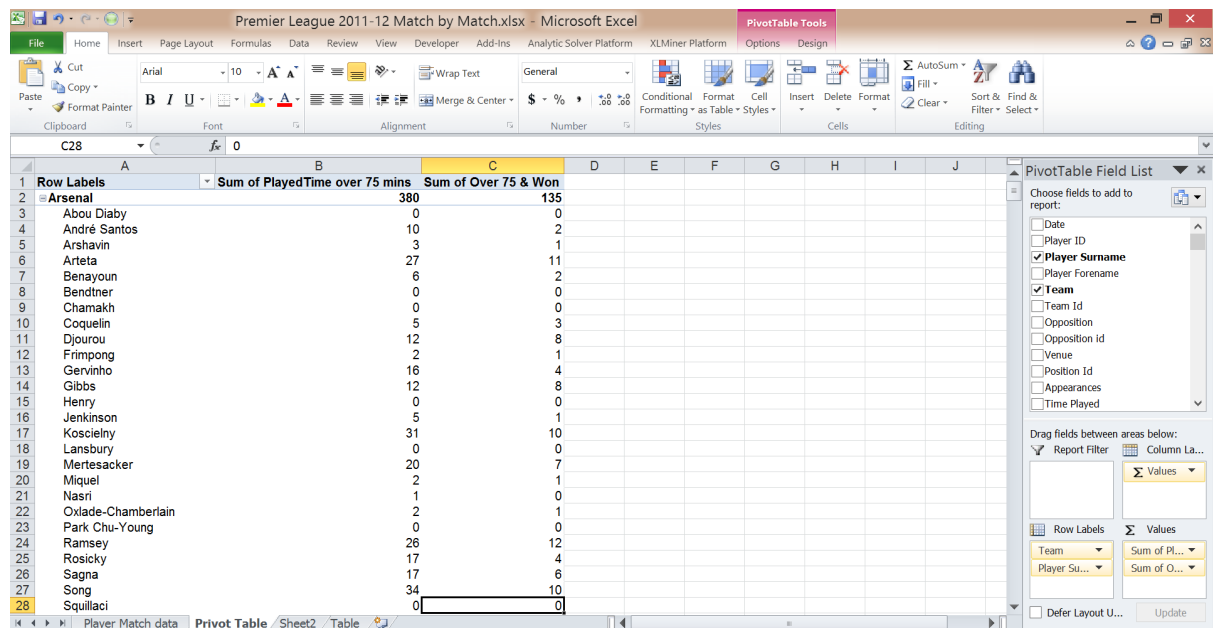


Figure 5: Pivot Table of players by Team using Sum of Played Time over 75 mins and Sum of Over 75 & won

After obtaining the desired data, we inserted a pivot table, choosing relevant fields required in our analysis, such as Team, Player Name, Sum of playedTime over 75 mins and Sum of over 75 mins & won.

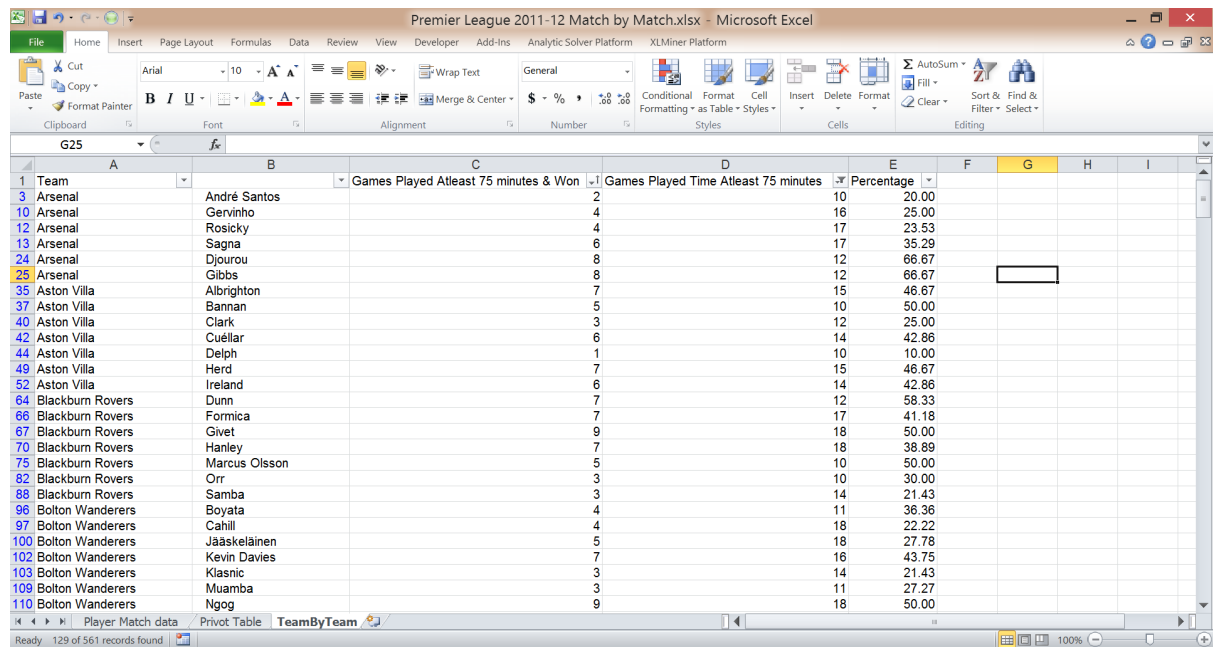


Figure 6: Players by Team and Percentage Win

We transferred these data into a new worksheet, TeamByTeam. Players who played 10 – 19 games in the season were filtered for us to calculate the percentage of the

win rate(Game played Atleast 75 minutes & Won/Games Played Time Atleast 75 minutes). The result was presented under the new column, Percentage.

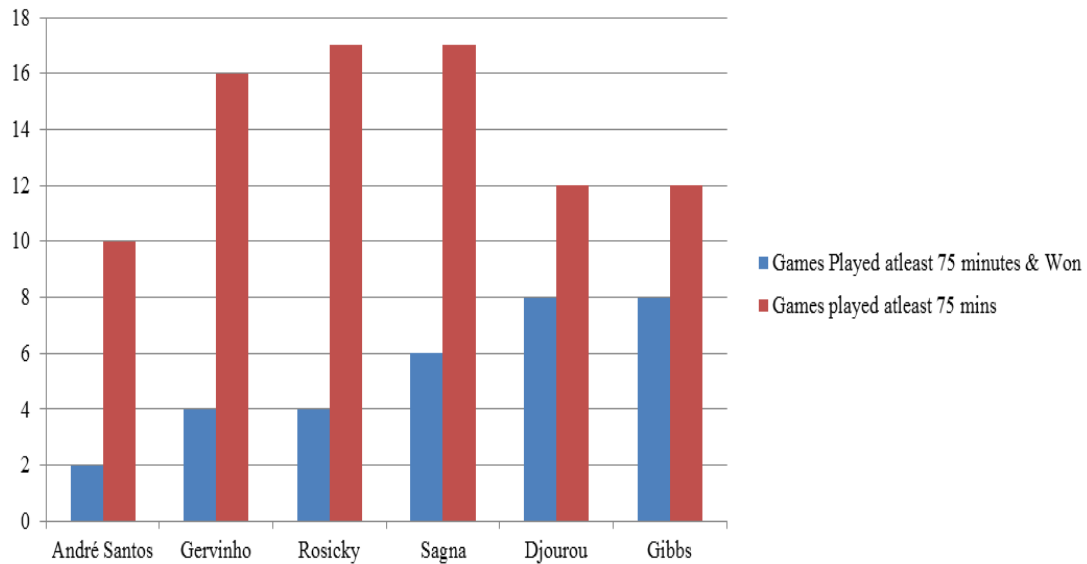


Chart 1: Games Played at least 75 mins & Won and Games Played at least 75 mins for Arsenal Players

Win Percentage

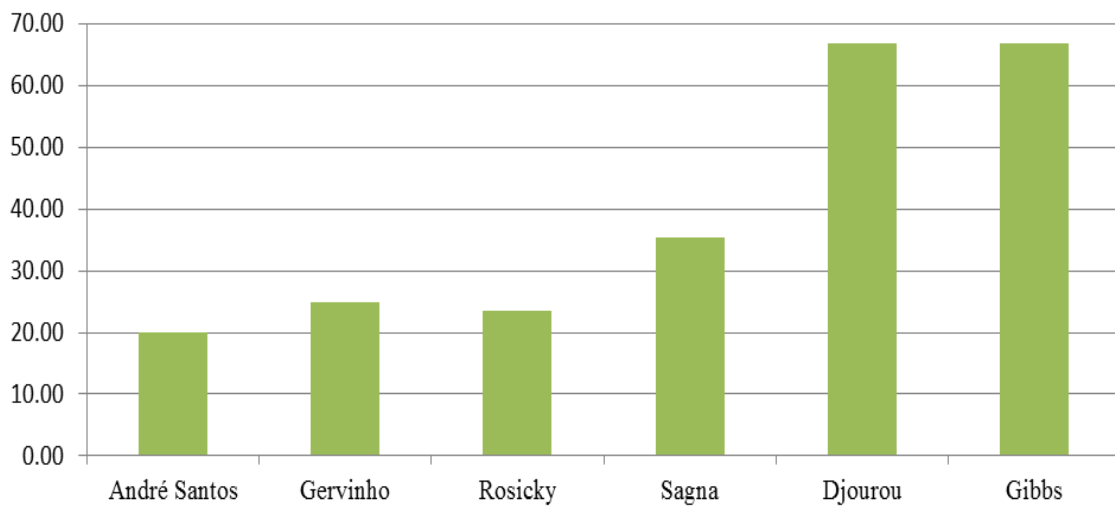


Chart 2 : Win Percentage of Arsenal Players

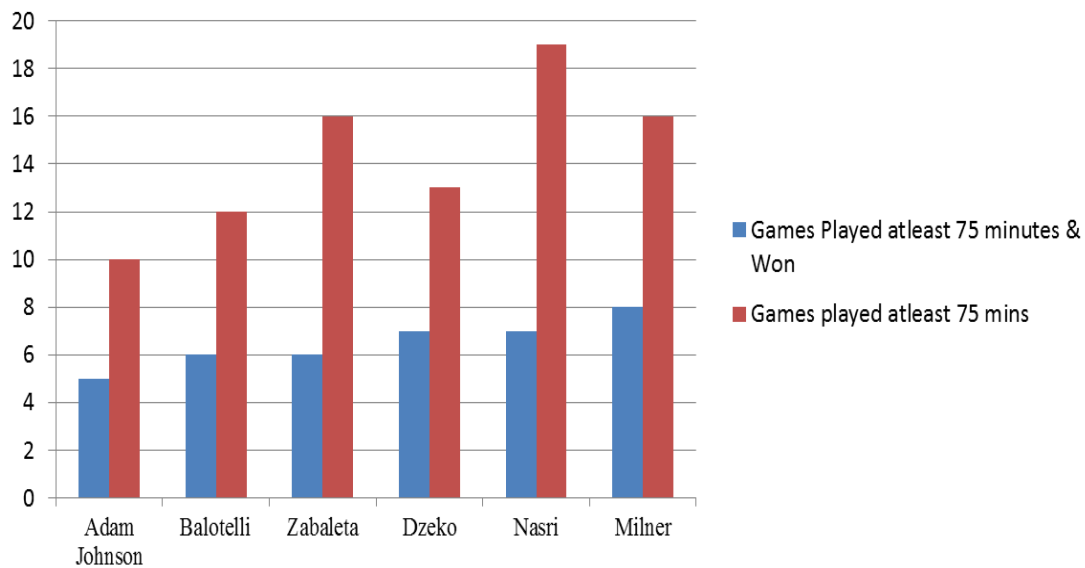


Chart 3: Games Played at least 75 mins & Won and Games Played at least 75 mins for MCFC players

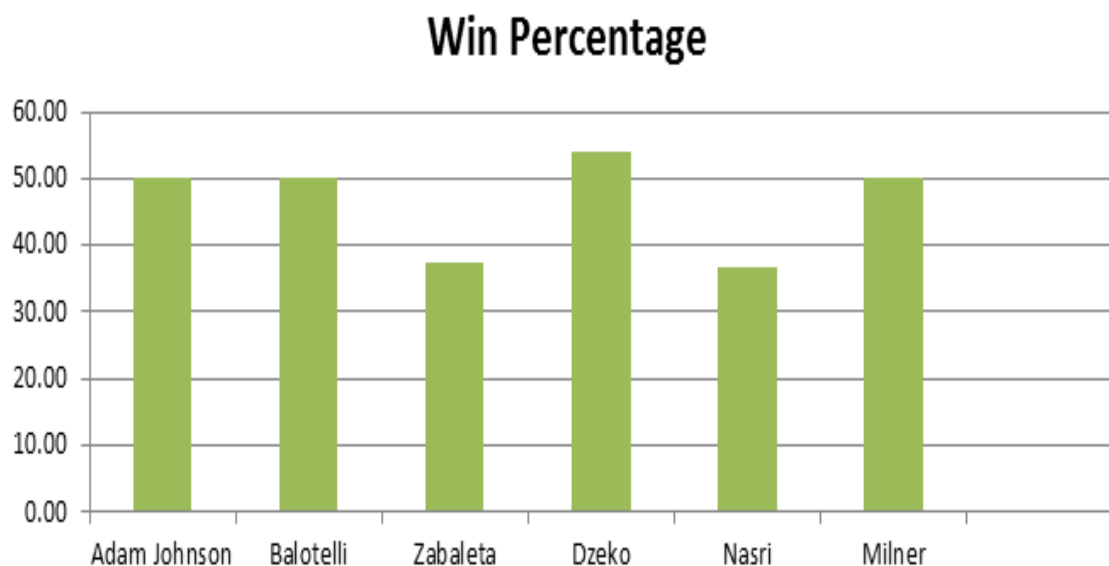


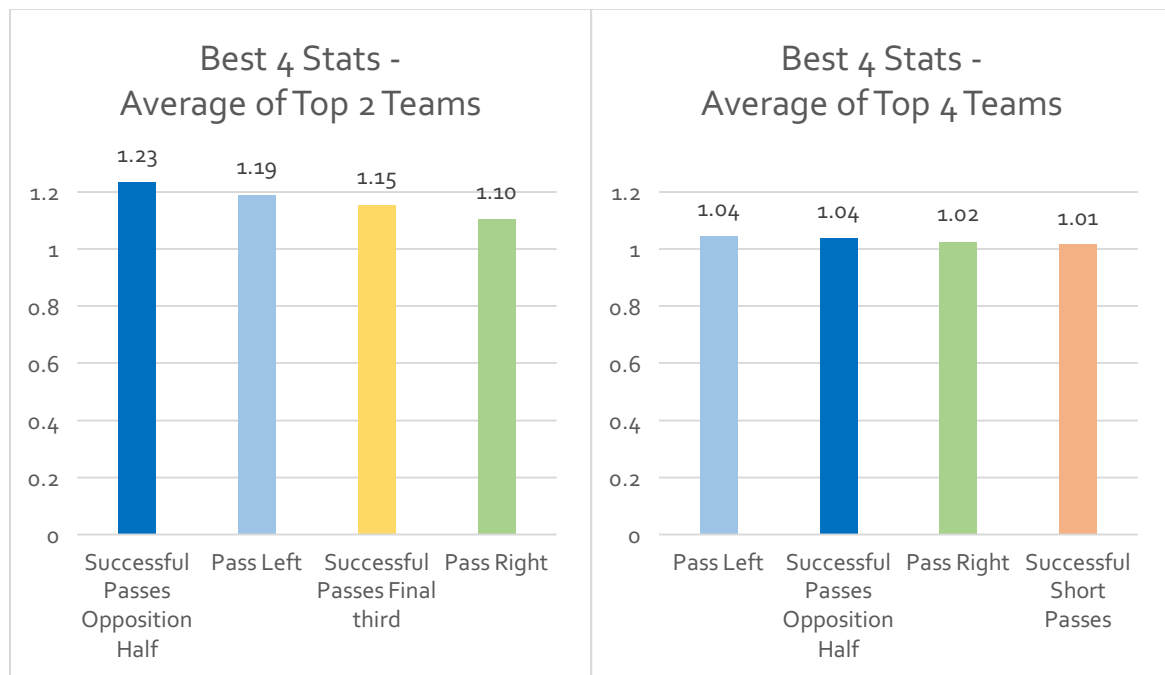
Chart 4: Win Percentage of MCFC Players

Lastly, we filtered out players who fit the requirement of playing at least 75 minutes. Using team Arsenal as an example, we can tell that the players Gibbs and Djourou had the highest winning rate of 66.67%. As for MCFC, player Dzeko had the highest win percentage, followed by Adam Johnson, Balotelli and Milner.

From our descriptive analysis on players who had good winning percentages, the team managers may be more well-informed in their selection of players for important football matches should they wish to maximize their chances of winning.

Descriptive Analytics

Best 4 Statistics of Top 2 Teams and Top 4 Teams



Charts 5 & 6: Best 4 Stats of average Top 2 teams and Best 4 Stats of average Top 4 Teams

We achieved these scores by using the Excel's Pivot Table to sort the match statistics, then normalizing them (such that the mean is zero and standard deviation is one), assuming the data follows a normal distribution curve. We then averaged the match statistic for the Top 2 Teams and Top 4 Teams. From this data, we can determine the best 4 statistics of the average of the top 2 teams and the average of the top 4 teams.

One of the patterns we observed is that the best statistics of higher ranking teams, among all others (less those related to shots), are passes. More specifically, Successful Passes in the Opposition Half and Passes to the Left.

There is a good possibility that the association in passes in the opposition half and passing to the flank and achieving a high ranking on the premier league board, is higher compared to the other statistics.

Top 3 attributes of Top 2 Teams

Manchester City	Success rate of successful passes in final third	Sum of Successful Passes Final third	Sum of Successful Lay-Offs
	1.404898469	0.997712832	0.863994388
Manchester United	Sum of Successful Passes Final third	Sum of Successful Lay-Offs	Success rate of successful passes in final third
	2.008366966	1.463611944	1.322190622

Table 1: Top 3 Attributes of Top 2 Teams

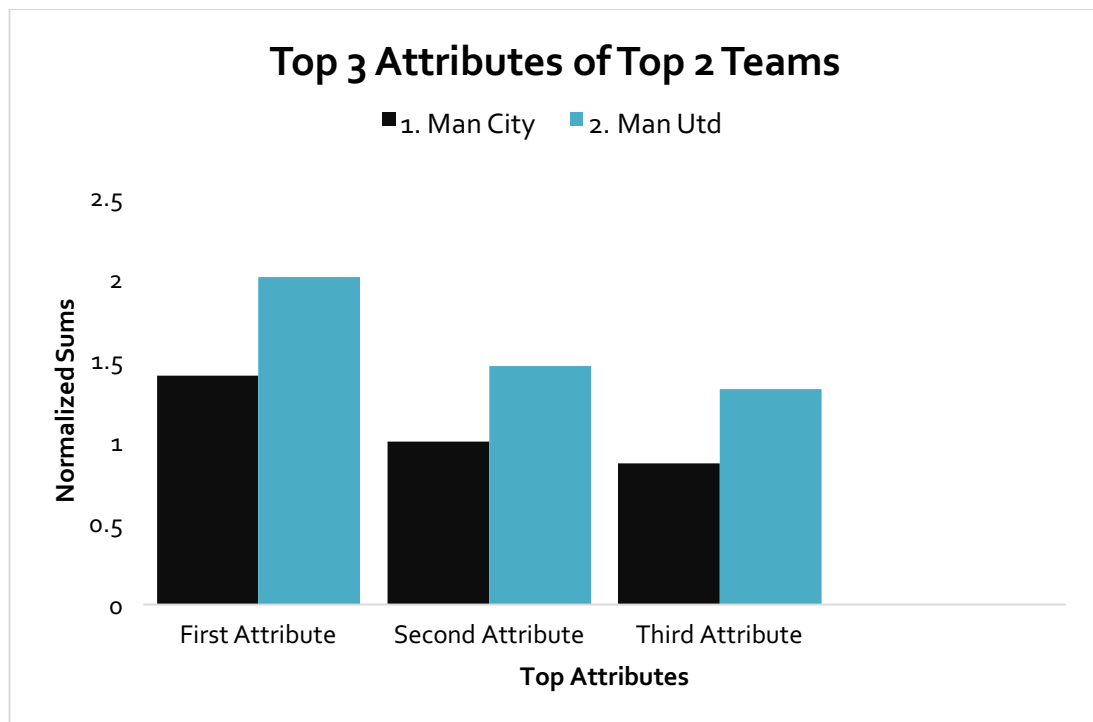


Chart 7: Top 3 Attributes of Manchester City and Manchester United

The objective for this descriptive analysis is to filter out the best qualities from top 2 teams, Manchester City and Manchester United in Premier League 2011-12. This time, we normalized the statistics at a team level instead.

These 2 winning teams have the high success rate of successful passes in final third, sum of successful passes final third and sum of successful lay-offs. From the data, these 2 teams performed above the League Average of 0 among other teams in the said Attributes. Hence, it is highly likely that the three attributes are important traits in determining the performance of the teams.

Predictive Analytics (Regression)

Ratio of Short/Long Passes =

$$10.42 + 0.96 (\text{Sum of Through Ball}) - 0.19 (\text{Sum of Offsides})$$

Dependent Variable	Independent Variable
Ratio of Short/Long Passes	Sum of Through Ball
	Sum of Offsides
Y-Intercept	
10.42	

Our regression process

1. We constructed a model with all available independent variables. We checked for the significance of the independent variables by examining the p-values.
2. We identified the independent variable having the largest p-value that exceeds the chosen level of significance 5%.
3. We removed the variable identified in step 2 from the model and evaluated adjusted R square, removing only one variable at a time. This approach allowed us to seek the highest adjusted R square value of 24.8% with our model shown above.
4. We continued until all the variables are significant. (When the p-value is lower than our chosen level of 5%, we will reject the null hypothesis and conclude that the slope, coefficient for each independent variable is not zero, which meant it is a statistical significant variable.)

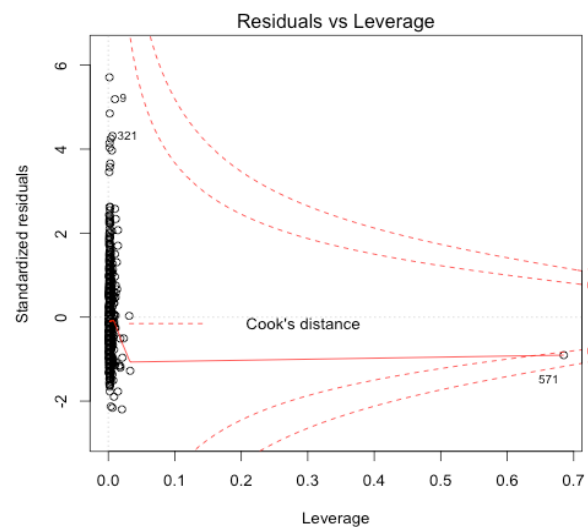
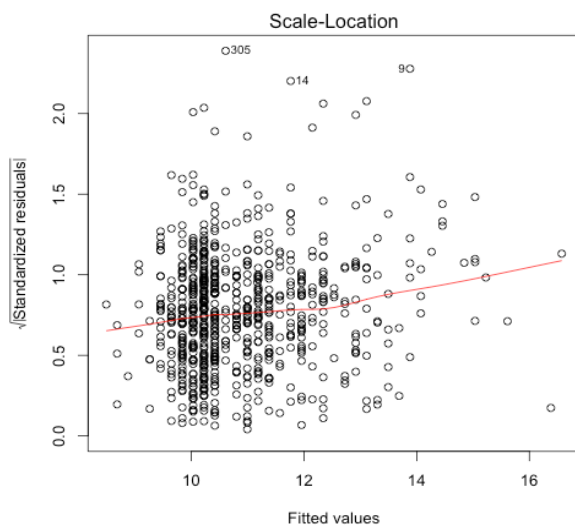
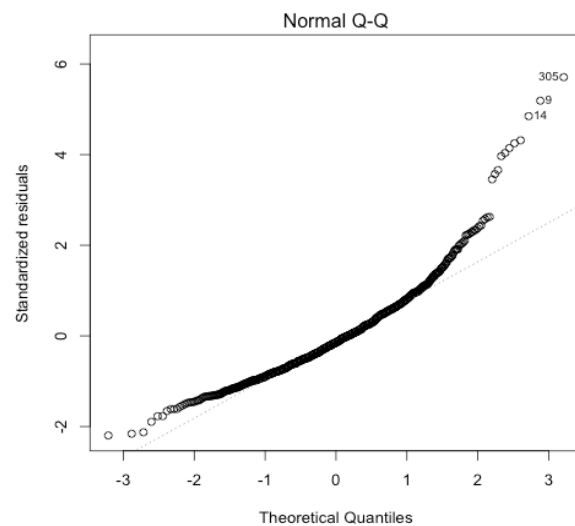
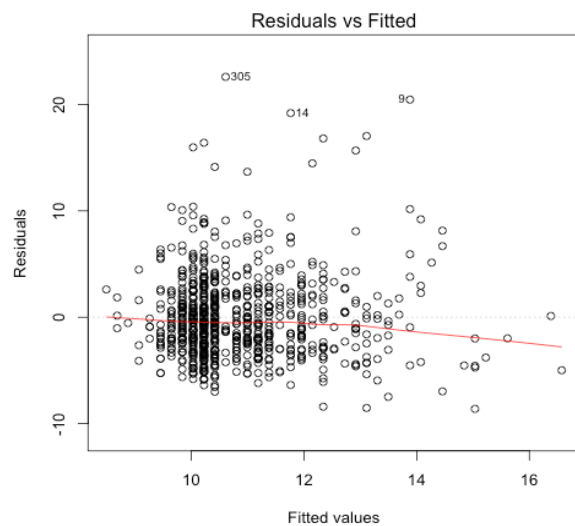
Our model followed the age old principle of parsimony with 2 independent variables and we removed the independent variables that are multi-collinear based of VIF values smaller than 10.

Linear Regression							
Regression Statistics							
R		0.28457					
R-square		0.08098					
Adjusted R-square		0.07855					
S		3.96052					
N		760					
Ratio of short/long passes = 10.41615 + 0.9624 * Sum of Through Ball - 0.19267 * Sum of Offsides							
ANOVA							
	d.f.	SS	MS	F	p-level		
Regression	2	1,046.25578	523.12789	33.35061	0.		
Residual	757	11,874.08147	15.68571				
Total	759	12,920.33725					
	Coefficient	Standard Error	LCL	UCL	t Stat	p-level	H0 (5%)
Intercept	10.41615	0.18808	10.04692	10.78537	55.38143	0.	rejected
Sum of Through Ball	0.9624	0.11823	0.73031	1.19449	8.14039	0.	rejected
Sum of Offsides	-0.19267	0.0486	-0.28808	-0.09725	-3.96405	0.00008	rejected
T (5%)	1.9631						
LCL - Lower value of a reliable interval (LCL)							
UCL - Upper value of a reliable interval (UCL)							

Figure 7: Linear Regression model using Ratio of Short/Long passes as Dependent variable

Multi-linear regression assumptions

- Based on the plot of Residuals vs Fitted, there is no specific pattern observed as residuals are randomly scattered about zero. In addition, the red line is flat along zero, which means linearity assumption is met. However, larger predicted values are associated with larger error of residuals as seen above in the Residuals vs Fitted graph, with increasing values of residuals, the variance inevitably gets larger as well.
- Based on the normal QQ or the quantile-quantile plot, the bulk of the points lies on the diagonal line, which assumes that the errors for each independent variable is normally distributed with a mean of zero. This is further supported by our large sample size.



- Residuals that are not independent of each other will show significant patterns on a Residual graph. As shown above, no pattern of significance could be observed thus the required regression assumption of residual independence is validated.
- A majority of n residuals in the graph showed no serious differences in the spread of the data for different values of the sum of independent variables, particularly if the outlier is eliminated, therefore, the assumption of homoscedasticity has been somewhat met.
- For cross-sectional data, the assumption of independence of error is usually not a problem, especially since time is not included as one of our independent variables.

Interpretation of the results

One of the ways to measure the overall predictive accuracy of a multiple regression model is the R-square value. If the R-square value is 1.0, this means the model explains 100% of the variance and thereby producing a result with perfect predictive accuracy, which is not possible in the real world, but rather varies according to context. In our study of an extended and multilayered social phenomenon like soccer, our value of 0.08 might be considered acceptable. The low R-squared value in our model shows that even noisy, high-variability data can have a significant trend, due to low p-values, which maybe comparable to a plot of a high R square value data.

S, Standard Error of Regression, represents the average distance that the observed values fall from the regression line. As a general rule of thumb, S must be less than or equal to 2.5 to produce a sufficiently narrow 95% prediction interval. Hence our model which showed that about 95% of the observations should fall within plus or minus 3.96% of the fitted line, meaning an improvement in precision might be required for our model.

As the assumptions of regression are not violated, statistical inferences drawn from the hypothesis test will more likely be useful than not. From our descriptive analytics, what we have decided to look for in our predictive analytics was to prioritize what football fans find exciting or look for in a particular game of football. With the top two teams having passing as one of the core reasons to their league success, and in addition to the popularity of the footballing philosophy *tiki-taka*, popularized by F.C. Barcelona

and the Spanish national team during the years of 2008 – 2012 (Roden, 2014) of or in layman terms, possession football; we have decided to look into the particular attributes within a match that results in a more possession based game. The dependent variable we are looking at is the ratio of short to long passes, which roughly translates into how a team is more likely to play with a *tiki-taka* philosophy, which directly relates to how exciting a football match are to the fans (Ball, 2014).

With these guidelines in place, the football club directors, when appointing a coach, may want to set out a certain philosophy the team plays with, as it affects the amount of fans attracted to their stadium and this in turn, affects their revenue earned. Therefore, with the predictive model we have established after running regression, there are certain attributes which teams can focus on to establish the *tiki-taka* playing style, resulting in an attractive style of playing which essentially means more fans attracted and thus help garner more revenue.

Predictive Analytics (Data mining)

Logistic Regression is a regression model used when the dependent variable is categorical. By inserting the independent variables into the model, we are able to obtain the probability of ending up with a particular categorical result.

In our case, we aim to obtain a model that will determine what factors significantly increases or decreases the chances of winning or not winning a match.

By obtaining these factors and their coefficients, we can compare which statistic has a greater impact in improving the probability of winning or not. If they have not been performing well, they may look into these factors and determine if they played a big impact in the outcome. Football coaches will then be able to know what kind of training they should place more emphasis on.

```
> summary(log.mod3)

Call:
glm(formula = raw_match_data$Sum.of.Winning.Goal ~ raw_match_data$Sum.of.Total.Unsuccessful.Passes.Excl.Crosses.C
  orners +
  raw_match_data$Sum.of.Duels.won + raw_match_data$Sum.of.Touches,
  family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7833  -0.9305  -0.7401   1.2063   1.9320

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.6443414   0.7379601  -4.938 7.88e-07 ***
raw_match_data$Sum.of.Total.Unsuccessful.Passes.Excl.Crosses.Corners -0.0113910   0.0052584  -2.166  0.03029 *
raw_match_data$Sum.of.Duels.won    0.0252770   0.0091066   2.776  0.00551 **
raw_match_data$Sum.of.Touches      0.0045542   0.0007344   6.201 5.61e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 997.04  on 759  degrees of freedom
Residual deviance: 940.42  on 756  degrees of freedom
AIC: 948.42

Number of Fisher Scoring iterations: 4

> exp(coef(log.mod3))
                    (Intercept)
                    0.02613862
raw_match_data$Sum.of.Total.Unsuccessful.Passes.Excl.Crosses.Corners
                    0.98867363
                    raw_match_data$Sum.of.Duels.won
                    1.02559917
                    raw_match_data$Sum.of.Touches
                    1.00456457

> log(1.02559917)/log(1.00456457)
[1] 5.55028
```

Figure 8: Results of running Logistic Regression on R

In our logistic regression model, we identified three significant factors (p -values < 0.05); the Total number of Unsuccessful Passes Excluding Crosses and Corners in a match, the Total number of Duels won in a match, and the Total number of Touches in a match.

The exponential of the coefficients represents the increase in the odds of winning for every 1 unit of the corresponding variable while all other factors remain constant. This means that the odds of winning is 0.989 times less for every unsuccessful pass excluding crosses and corners made, under the condition that the other variables remain the same. From the exponential of the coefficients, we can also see that the Total number of Duels won increases the odds of winning more than the Total number of Touches. With some calculation (at the bottom of the screenshot), we can more specifically identify that one additional duel won is as important as making 5.5 additional touches in a match.

With this model we not only know the variables that are significant in their impact towards the probability of winning, but also which variables are more important so that the trainings can be more tailored to those factors.

Limitations for Logistic Regression

As seen from the summary of the logistic regression, the AIC value is quite high. This means that the model is not a very exact fit, even though it is significant. This could possibly have been overcome with more match data from other premier leagues as there were only 176 matches that are played in one premier league.

Conclusion

From our descriptive analysis, we have not only identified unobvious attributes to spot good players. We also identified that passes in each team is especially crucial in determining a team's performance in the League, and thus contribute to its win.

At the same time, logistic regression has also revealed that the Total number of Unsuccessful Passes Excluding Crosses would leads to lower odds of winning whereas the Total number of Duels won and the Total number of Touches would be increase the odds of winning. Hence we should prioritise the training of such attributes to strengthen a team and to reach a better league position.

We have also derived a linear regression model with specific attributes which could contribute to an exciting match. Based on the Tiki Taka mentality, this would be especially useful to the football manager should he intend to attract more crowd and fans. He could then focus on those specific attributes to achieve a greater revenue.

While a football manager could have been able to gain relatively useful insights from our analysis, there were also various limitations that we faced while implementing our models. One of which is that the model may not be an exceptionally good fit and therefore, it could be improved by including more varying independent variables e.g. x^2 , $(x_1 + x_2)$ into our regression model. With respect to the data we have used, football matches can be very unpredictable. Hence, more specific and useful data such as the success rates of in swinging and out swinging corners could improve the reliability and accuracy of our analysis.

References

- Ball, P. (2014, November 18 November 2014). *FC Barcelona: The science behind their success*. Retrieved from BBC: <http://www.bbc.com/future/story/20141024-what-makes-barcelona-so-special>
- ESPN FC. (n.d.). *Barclays Premier League Table - ESPN FC*. Retrieved from ESPN FC: <http://www.espnfc.com/barclays-premier-league/23/table?season=2011>
- Roden, L. (2014, June 13). *Drop the 'tiki taka' myth: Spain's football is Dutch*. Retrieved from talkSPORT: <http://talksport.com/football/drop-tiki-taka-myth-spains-football-dutch-14061396140>