**Statistics Final Project**

**EXECUTIVE SUMMARY**

The main purpose of our project is to analyze soccer games as well as top soccer leagues across Europe and to find interesting research questions that can better our understanding about the game on a country, league and team level. In order answer our research questions, we divided our analysis into three parts.

In the first part, we used data set related to the match itself to build regression model for soccer games and analyze different style for different league. For the regression model for England, Germany and Spain, the significant variables include home shoot target, away shoot target etc., and there is mediation effect from shoot on target when corner influence on the game result. For France model, we found shoot target has mediation effect on shoot and game result. For Italy model, we found its unique defense style.

The second part of our project is to apply our outputs in the first section and conduct team based analyses. The main objective is to find out what are the best teams across all five major leagues in the last five years. The objective criteria has been separated into two sections: offense and defense. In order to accomplishment the objective, we did a series of hierarchical clustering analyses to filter teams with better performances. In the offense section, we find out that English Premier League and Spanish La Liga are the strongest leagues with 19 and 13 team appearances. In the defense section, we find out that French Ligue 1 and Italian Serie A are the strongest leagues with 24 and 17 team appearances.

The third part of our project looked into the significant elements of the Top 5 teams in the EPL to see if there were any commonalities that made them so successful, or any significant differences that made them unique. Our research found that one common factor between the teams was the fact that they were able to “Come Back” from being down in a game more effectively and be able to secure a Win more often than other teams. Differentiating elements that led to success included for example aggressive gameplay and fouling from Manchester United, free kicks and corners from Liverpool and counterattack play from Arsenal.

Looking beyond the important factors of championship teams, we wanted to see if there is any significance to predicting wins based on oddsmakers predictions. Using betting data from BET365, an analysis was conducted which found that the inefficiencies in oddsmakers predictions were only found in games with significant underdogs i.e. probability of winning set at less than 15%. Looking further into underdog games, it was found that when the significant underdog was playing at Home the probability of them winning overshot oddsmakers predictions by nearly 30% in some cases, providing an opportunity for betmakers to make significant profits when focusing only on underdog home games, for the most part because of “Home Team Advantage”

**PART 1**

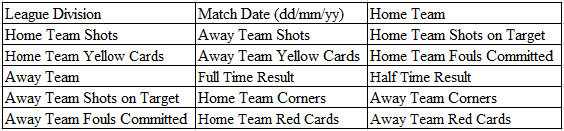
Game Based Research Questions

1. What are the variables that are considered significant in predicting a win in a game of soccer for the past 10 years?
2. How do these variables change/ transform when looking at Wins in 5 soccer dependant nations? What does this tell us about each nation and leagues style of play?

**Data Set**

Our data set is from<http://www.football-data.co.uk/data.php>. It provides Historical Football Results and Betting Odds Data for top leagues all over the world from season 1993/1994 to season 2015/2016. The data sets we used are league football matches data of English Premiership, Italian Serie A, German Bundesliga, Spanish Primera Division and Championnat de France de football Ligue 1, which are known as the top 5 soccer leagues in the world, from season 2005/2006 to season 2014/2015. The data set includes 62 columns.

For the first part, we will use columns related to the match itself to build regression model for soccer games and analyze different style for different league. The total row number for the 5 leagues is 17,000. Below are the terms of the columns:

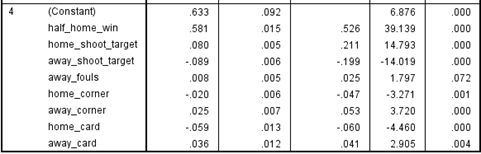


**Data Transformation**

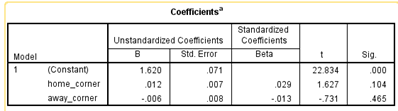
Since there are two teams for each match, we will stand on home team side to build the model. The bold columns were transformed for modeling. The full time result was transformed to home score: 3 for win, 1 for draw and 0 for lost. For half time result, we did the same transformation. We use the home card column to instead home team red card and yellow card: 1 red card = 2 yellow card. For away team card result, we did the same transformation.

Then we used backward method to build linear regression for each league.

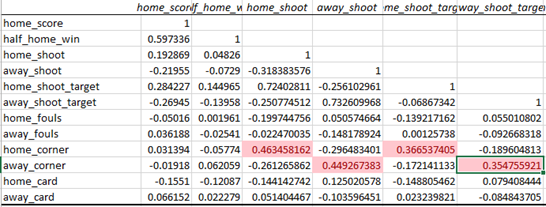
**England**



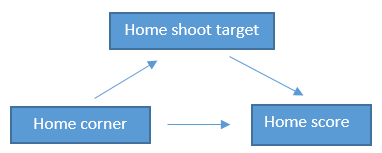
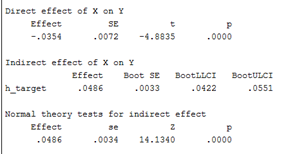
From the coefficients of home and away corner we could conclude that more home corner and less away corner will lead to lower home score, which seems doesn't make sense. If we just included home corner and away corner in the model, we got opposite coefficients. So the coefficients of corner are affected by some other variables in this model.



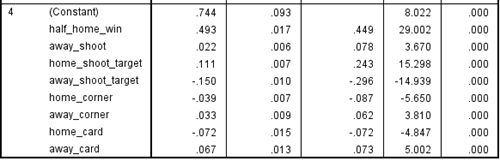
From correlation table, we could see that home corner and away corner have high correlation with shoot and shoot target.



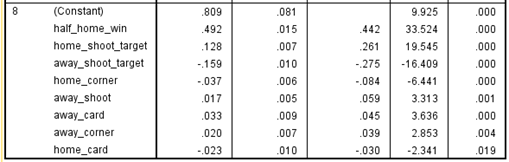
There is partial mediation for home shoot target on home corner to home score. If home corner could become shoot target, it will bring a higher score for the home team.



**Germany**

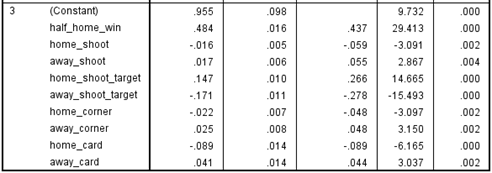


**Spain**

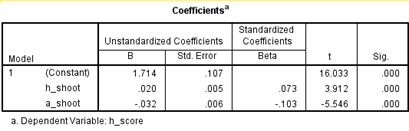


The models for Germany and Spain both got negative home corner coefficient and positive away corner coefficient, which is similar to England. We won’t repeat it in the report.

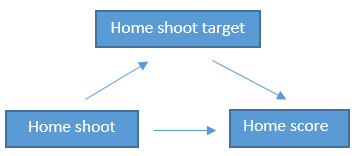
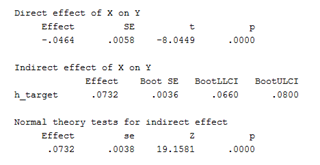
**France**



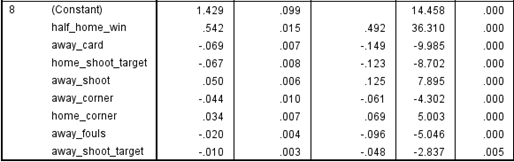
We noticed home\_shoot is negative. If we just include home shoot and away shoot in the model, the coefficients are opposite.



There is mediation for home shoot target on the effect of home shoot to home score. Same as home corner, shoot on target could make shoot more meaningful for the final score.

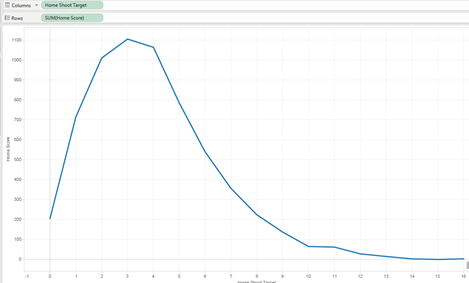


**Italy**

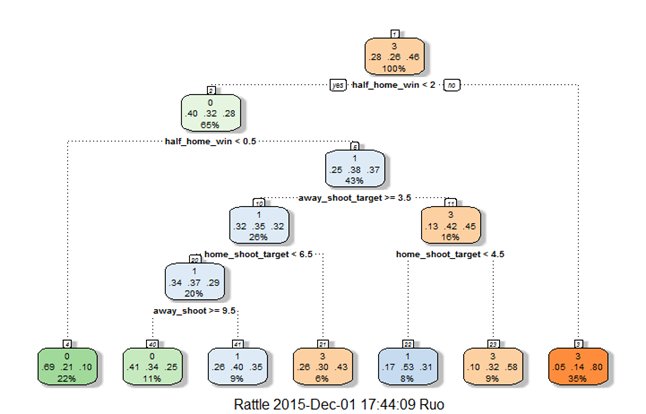


Different from those four leagues above, for Italy, away card coefficient is negative. If we just use away card to run the model, the coefficient is still negative (-0.128, the effect is stronger than the model with all of variables), which means more yellow or red card for away team means lower score for home team. Away fouls gave us same information. So we can refer that more “effective” fouls may lead to higher score in Italy league.

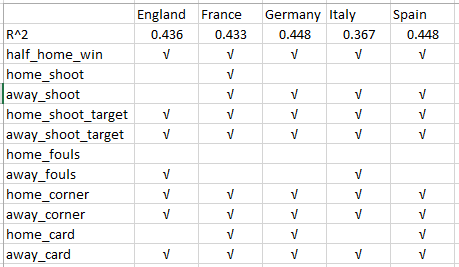
Another finding is home shoot target is negative here. We run another model just with home shoot target, the coefficient is -0.084. The relationship between home shoot target and home score is showed below. When home shoot target is less than or equal to 3, it has positive effect on home score, but when it keeps increasing, home score becomes decreasing. We guess it could be the goalkeeper’s credit. Less but effective attack and tight defense is always Italy style. Besides, the analysis about fouls above could be another evidence for Italy team's’ defense. We will do further analysis about it in part 2.



**Analysis for all tables combined**



We used regression tree for the model with total data. The most important IV for predicting the final result is half time score. If the home team is in the lead before the intermission, it has 35% chance to win the game. If the home team is in behind before the intermission, there is 22% possibility for it to lose the game. If the game is a draw before the intermission, the second important IV is away shoot target. If away shoot target is less than 3.5, the home team has 17% chance to get win or draw. If away shoot target is greater than 3.5 and home shoot target is greater than 6.5, home team has 6% chance to win. If away shoot target is greater than 3.5 and home shoot target is less than 6.5, the game result depends on away shoot number. If away shoot number is greater than 9.5, there is 11% chance for home team to loss.



Above table shows model comparison among all five leagues. There are some common variables in those model (half\_home\_win, home\_shoot\_target, away\_shoot\_target, home\_corner, away\_corner,away\_card), which shows similarity for those league and we will use them for further analysis. At the same time, we noticed there are some variables, such as home\_shoot, were significant for some of models, which gave us the information that there is difference among those leagues. In the next part, we will use those variable for further analysis about different style for each league.

**PART 2**

After doing the game based analyses, we applied the results and came up with the team based research questions:

1. What are the variables that make great teams?
2. Which are the teams and leagues that a person who wants to start watching soccer should follow to maximize their enjoyment?

In order to find out, we did multiple hierarchical cluster analyses in R to choose the best teams. The tables we used contain all the factors that measure team's performance from the five leagues in the last five years. We separated the analysis into two sections: offense and defense and we picked appropriate variables that have shown their significance in previous regression analyses.

Data Source: <http://www.footstats.co.uk/index.cfm?task=Leagues>

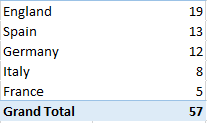
**Offense**

In this section, the first variable we used is “perGD”, which is goal differentiation between one team and the against team for each game on average. For example, if one team made 80 goals and allowed 40 goals in one season and there were 20 games in that season. The “perGD” index would be (80-40)/20=2. This variable tells us how dominate teams are in each game.

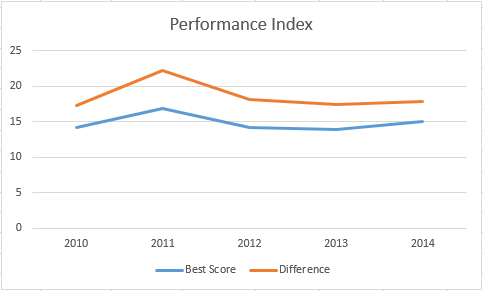
Second variable we used was “perSht” (shoots on target for each game on average). This is the most intuitive offensive measure in a soccer game. Also, both of these two variables are significant in the previous regression analyses. We then created a new column called “score” that is the multiplication of “perGD” and “perSht”. This index helped us conduct the hierarchical cluster analyses since we want teams that with both of these two variables as high as possible. We did five cluster analyses and each analysis contains all the teams in that year. Furthermore, we believe there should be 7 clusters for each analysis based on the size of the data. We picked the clusters with highest “score” number and made a pivot table of the selected teams as well as appearance count that is shown below:



Based on the result we can see that Manchester City, Real Madrid, Barcelona, and Chelsea have shown the best offensive performance with full appearances in the last five years. Manchester United, Arsenal, and Dortmund are in the second tier with one absence. Milan, Bayern Munich, and Leverkusen are in the third tier with two absence. This result does not surprise us a lot because most of the clubs above are extremely popular and famous in soccer world and they can always gain top standings within their own leagues. Leverkusen might be the biggest surprise in this result because there are few more candidates in German Bundesliga other than Bayern Munich and Dortmund to make this spot such as Schalke 04.



As for leagues, English Premier League has the strong performance with 19 appearances. Thanks to Real Madrid and Barcelona, Spanish La Liga locks second seat. German Bundesliga is in the third place with 12 appearances. The result of Italy and France show that they are not as offensive as other leagues even though they also have very well-known and exciting clubs like Milan, Juventus, and Paris Saint Germain. This output also matches our common sense and intuition. English Premier League teams and matches are known for their competitiveness and enjoyment of watching.



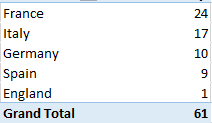
The graph above shows the team's offense performance index (“perGD” \* “perSht”) for the last five years. The blue line represents the clusters with highest “score” number, orange line represents the difference between the clusters with highest “score” number and the clusters with lowest “score” number. These two lines follows the same trend and the overall trend is pretty stable except the peak in 2011 to 2012 season.

**Defense**

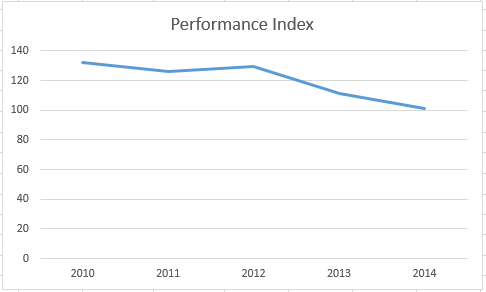
In this section, We first created a column called “perGameAllowed” (one over goals allowed in one season divided by games played). This column tells me which team loses least goals per game on average but because we used 1 over the original calculation. A higher number will show a better defense performance. Then we choose to use “Foul.F”, which is a measure of how many fouls that one team made for each game on average. A higher number of “Foul.F” means this team could either be very competitive or very dirty. We pick this measure because we want to filter teams that are not competitive. We also added another column that is called “foulEff” (“Foul.F” divided by how many yellow and red cards per game, which 1 red card counted as 2 yellow cards). This tells us how effective and tricky team players are because they know how to make non-serious fouls that would not be punished. Lastly, I multiplied “perGameAllowed” by “Foul.F” and “foulEff” to create this new defense index called “noscore”. A higher number of “noscore” means better defense. Same kind of analysis is performed and below are the selected teams as well as appearance count with highest “noscore” number:



Based on the result we can see that no team appeared in all of the past five years meaning it is harder to maintain a defense status than offense. Bayern Munich and Juventus are the best defense teams with four appearances. Monaco, Milan, St. Etienne, Toulouse, Atletico Madrid, and Lyon are in the second tier with three appearances. There are no team with three appearances of defense has more than two appearances of offense meaning it is exceptionally difficult to be good at both offense and defense. Bayern Munich is the only team that has at least three appearance in two categories showing its consistency and dominance in Bundesliga.



After reviewing the league's’ result, the first thing we noticed is that this defense table has a reversed order in compare to the offense table. This means that no league is equally strong in both offense and defense. France and Italy now lock the first two seats and England only has one appearance (Chelsea saved the country). This result somewhat matches our impression because Italian clubs are always known to be good at defense. French result is a little surprising because there are not a lot of public reports illustrating their typical game and team style. From the previous table we can also see that even though they don’t have teams that take over the first tier, there are plenty of French clubs with three or two appearances. French Ligue 1 deserve more attention based on their defense performance.



Similar to the previous analysis, the graph above shows the team's defensive performance index (“perGameAllowed” \* “Foul.F” \* “foulEff”) for the last five years. The trend line is showing team's defensive performance with the highest “noscore” number and it tells us that there is a gradual decrease. This could mean that teams are less focus on defense than offense, or it could mean that soccer is getting less physical than before.

**PART 3**

After completing the team based analyses, we are confident to conclude that English Premier League is the most exciting league to watch and the most competitive league across five major leagues. We want to take our research forward and discover more about the championship teams and research the predictions made by oddsmakers in the EPL:

1. What are the variables that make the winning teams in the EPL so prolific? Do these variables differ by team?
2. What do the odds set by oddsmakers tell us about the probability of a given team winning? Are there any inefficiencies set in these odds that can be leveraged?

Data Source:<http://www.football-data.co.uk/data.php>

A backward regression for the Top 5 winning teams in the English Premier League was conducted to see if there are any significant variables that differentiate these teams from each other as well as of the rest of the league. An analysis of these variables should be able to tell us in greater detail if there are common factors that link these “Winning” teams which make them as successful as they are or if each team is uniquely successful in its own way and form, distinguishing it from the rest.

From this analysis we hope to gain better insight into what differentiates the most prolific teams in the EPL and how a given viewer can choose which team to follow based on his interests.



1. Arsenal

From the significant factors Away team shots and Away team fouls seem to have a slightly positive coefficient which is unusual but it means a high action game with more free kicks and quick gameplay benefits Arsenals offense. Away goals at half time is also beneficial suggesting Arsenal tends to win games when they are down at half time, however they tend to lose a higher percent of games when they are up at half time. Half time points being significant suggests that one can predict the final outcome of an Arsenal game based on their position at half time. Full time home goals is positive and away goals negative which makes sense.

2. Manchester United  
  
Home team red cards is a significant predictor of Manchester United winning the game. This suggests that the more aggressive Man United is the more likely they will get the win despite red cards- this could be an allusion to their nickname “The Red Devils”. Half time home goals has a negative coefficient suggesting Manchester United maybe better in the second half in getting the win or losing their goal advantage.

3. Manchester City

A lot more significant variables for Manchester City compared to others, this may be because they are new to the championship category of the EPL, hence a higher focus on all aspects of gameplay. Away team corners are positive-which means they are good at defending corners and allowing corners could be a waste of time for the opposition-however it’s not highly significant. First team where Home shots on target is a significant predictor which makes sense which ensures an exciting and offensive gameplay. Away goals is a good predictor of a win or loss however home team goals is inverse, suggesting a lead at half time does not guarantee a full time win. Home team yellows is also positive and significant suggesting Manchester City has an aggressive style of play that could be a good predictor of wins.

4. Chelsea

Shots on target are significant and positive which suggests an offensive style of play that maximizes on shots taken within the borders of the goa. Once again Half time Away goals are positive suggesting that Chelsea like all other winning EPL teams is considered a “Comeback” team. This maybe the one commonality that can be seen for all the EPL championship teams which makes following them so exciting.

5. Liverpool

Away red cards benefit Liverpool making it easier to carry out its attack, suggesting a good percentage of its attacking play are based on a weakened Away defense. Negative Away shots on target suggests that Liverpool’s defense and goalkeeper are not upto par with the rest of the championship teams. Half time home goals have a small negative coefficient suggesting Liverpool too is a comeback team like the others. However, Away fouls are positive suggesting Liverpool is offensively dangerous in fixed spot plays such as freekicks and corners. It is also hard to predict a Liverpool win based on half time stats.

**Inefficiencies and Odds**

So we know that EPL is the most exciting league with high offense and almost low defensive factors amongst championship teams. The regression analysis above also tells us that the single commonality between the championship teams is defined by their “Comeback Ability” and the fact that they will walk away with a Win more so than others.

We answered the question on which League to follow for maximum entertainment and researched the championship teams in this league to try and find out why they are so prolific but now we want to use the betting dataset that we have to try and predict which games/ teams to bet on within the EPL to try and maximize our monetary outcome and add the element of “Risk” in enjoying a soccer match.

In general, betting companies and oddsmakers have highly sophisticated systems of producing odds- While they may not be perfect, it tends to be tough to find any consistent exploitable inefficiencies. In other words, it is rare that the odds of "Liverpool winning at home", or some other event like that, are consistently over or underestimated. You may think that the odds in an individual game may be incorrect, but in the long run inefficiencies like that rarely persist.

Why? Because bookies would lose money on them. If they realize they are starting to lose money, the odds are going to be adjusted to better reflect the probability of each result occurring. Additionally there are several complex algorithms and experts opinions that are involved in making a finals odds decision including qualitative factors such as team motivations and injuries.

Yet at the same time odds provide a good basis of being able to predict a win. For example, Bet365 put the odds of an Arsenal victory versus Chelsea at 2.38. These decimal odds imply that they expect the probability of an Arsenal victory to be about 42%.  The odds for relatively evenly matched games like the one above are probably pretty accurate, or at least more accurate than your average person- this is because of the high amount of viewership and actual bets being made during major games.

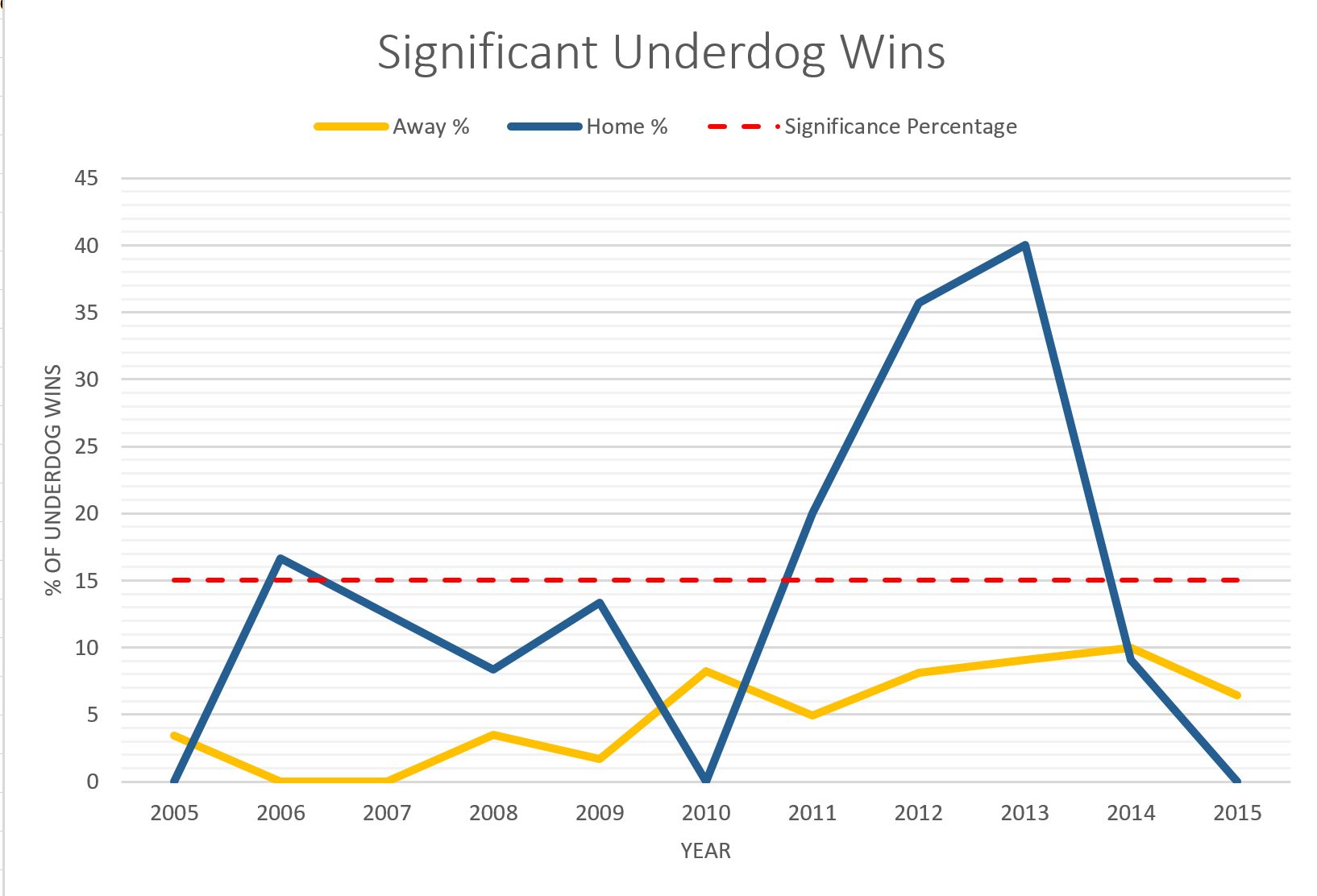
But what about significant underdogs? What about Manchester City(#1) against Cardiff(#19)? These are a little more difficult to assess. It's clear that Cardiff is an underdog in this game, but how much of an underdog? And do oddsmakers do a good job of assigning implied probabilities to these lopsided games?

R Studio was used to calculate the odds probabilities with the following steps. A cutoff of 0->15% chance of winning was assigned to define an Underdog team. Then a comparison was carried out to see how many times an underdog outcome actually occurred for both Home and Away underdogs.

Steps:

1. Create relevant percentage variables
2. Find and split data based on underdog games i.e. where home/away/ draw odds are set less than 15% of the time
3. Rows where away team was the underdog: 519
4. Rows where home team was an underdog: 88
5. Rows where draw is considered an underdog : 24

This makes sense because away teams are usually given lower odds and considered underdogs because of the significance of the “HOME TEAM ADVANTAGE” – considered highly significant in soccer where a home crowd can make or break games.



Significance of the Home team advantage is clearly evident when using Betting Data odds. The bookies aren’t always right!

The red dashed line is the 15% cutoff. For away underdogs, we see about what we would expect: the percentage of underdog outcomes occurring fluctuates, but it remains under the 15% line. For home underdogs though, there is a different story. For each year besides 2007-2010, home underdogs actually won more than 15% of the time, in some cases significantly more so. This seems to indicate that odds makers are underestimating home field advantage when weaker teams play stronger teams at home. As a bet maker one can take advantage of this and place bets on the underdogs where the odds may be low but the payout has the potential to be quite high.

This evidence seems to indicate that there is an inefficiency present, specifically in the odds of underdog home teams. However, there are still a few caveats that should be mentioned. First, there were a smaller number of home underdog games in the past 5 years, which may be influencing the results somewhat. Second, odds are made so that they somewhat lessen the payoffs so that the odds makers can make money. This means that if you find an inefficiency like the one above, it has to be large enough so that you can overcome the advantage that odds makers have when they make the odds. For the first 3 years, this does not seem to be the case for home underdogs. However 2011- 2014 would have been prime years for the smart betters.

While the trend in this line is currently negative, which could mean oddsmakers may have figured out such inefficiencies it may be wise to keep track of these odds and attempt to see if there maybe some seasonality or repetition in the time series graph which could result in a positive slope in the next 5-10 year period. This will provide a juicy opportunity for bet makers to make high profits.