Assessing the Accuracy of Betting Odds in European Football

Joseph Edward Pym

Student I.D.: 8404110

Supervisors:
Dr. Maha Moustafa and Dr. Alun Owen

A dissertation presented in the School of Computing, Electronics & Mathematics, Coventry University, for the degree of Bachelor of Science in Mathematics & Statistics.

Submitted April 2021

Acknowledgements

I would firstly like to thank my supervisors, Maha and Alun, for their help and expertise throughout the year: I couldn't have asked for any more of them. Thank you.

More thanks will go here.

Thank you!

0.1 *Abstract*

In this paper, we look at the accuracy of the odds offered by bookmakers on European football matches from the 2005/06 season, to the 2019/20 season, using a range of methods, inlcuding visual analysis, creating predictive models with correlation analysis, and finally, by using our findings to propose a new method of placing bets.

Contents

0.1 *Abstract* . . . 2
0.2 *Disclaimer* . . . 8

1 Introduction **9**
1.1 Background Information . . . 9
1.1.1 Association Football in Europe . . . 9
1.1.2 Probabilities, Odds & Gambling . . . 10
1.1.3 Betting Markets of Interest . . . 11
1.2 Literature Review . . . 11
1.3 Rationale, Aims, Objectives and Methods . . . 11
1.3.1 R Packages Used . . . 12
1.3.2 The Data . . . 13
1.4 Structure . . . 14

2 *Elite* European Leagues, 2005-20 **16**
2.1 Initial Data Analysis . . . 16
2.2 Exploratory Data Analysis . . . 19
2.3 Correlation Analysis, and Model Creation . . . 26
2.4 Measuring Predictive Performance . . . 28
2.5 Comparing leagues . . . 29
2.6 The effect of competitive balance on bookmaker accuracy . . . 30
2.6.1 What is competitive balance? . . . 30
2.6.2 Quantifying Competitive Balance . . . 30
2.7 Comparing Seasons . . . 32
2.8 Principal Component Analysis . . . 33
2.8.1 By-League Principal Component Analysis . . . 33
2.8.2 By-Season Principal Component Analysis . . . 35
2.9 Conclusion . . . 37

3 English & Scottish Leagues, 2005-20 **38**
3.1 Exploratory Data Analysis . . . 38
3.2 Correlation Analysis . . . 46
3.3 Comparing Levels . . . 48
3.4 The Overround . . . 51
3.5 Conclusion . . . 56

4 A Proposed Betting Method **57**
4.1 The Method . . . 57
4.1.1 Alternate Method . . . 57
4.2 Results . . . 57
4.3 Comparison . . . 57
4.4 Conclusion . . . 57

5 Conclusion **58**
5.1 Our Findings . . . 58
5.2 Challenges . . . 58

Appendices **63**

A Definitions **64**
A.1 Mathematical and Statistical . . . 64
A.2 Gambling Terms . . . 65
A.3 Acronyms . . . 66

B Chapter 2 Code **67**

C Chapter 3 Code **80**

D Chapter 4 Code **96**

E Project Diary **110**

F Word Count **115**

G Ethical Approval Certificate **116**

List of Figures

2.1 Histograms of the consensus probabilities for each outcome, in the French Ligue Une for the 2016/17 season. N.B. the change of scale on the y axes. 18
2.2 Boxplots of the consensus probabilities offered for each outcome. 21
2.3 Density plots for consenus probabilities offered for each outcome in the 1X2 market. 22
2.4 Density plots of the consensus probabilities offered for each outcome of the 1X2 market, split by league. 24
2.5 Tile Plot of the Correct Consensus Probability for each possible result 25
2.6 Scatter plot of the linear models created. 27
2.7 The variation of R^2, RMSE, P_1 and P_2 over time. 33
2.8 By-League PCA figures. 35
2.9 By-Season PCA figures. 36

3.1 Density plots of the consensus probabilities offered in the 1X2, UO and AH markets. 41
3.2 The AH Handicap versus the consensus probability of a Home Win, split by level. 42
3.3 The AH Handicap versus the consensus probabilities of both a Home Win and Away Win. 43
3.4 Tile plot of the correct 1X2 probability versus the full time result. 43
3.5 Tile plot of the correct UO probability versus the full time result. 44
3.6 Count plot of the home handicap offered versus the consensus probability of over 2.5 goals. 45
3.7 The actual v. observed goal difference. 45
3.8 Consensus v. Observed probabilities on the English & Scottish data, 1X2 market. 47
3.9 Consensus v. Observed probabilities on the English & Scottish data, Under/Over 2.5 Goals market. 47
3.10 Consensus v. Observed probabilities on the English & Scottish data, Asian Handicap market. 48
3.11 Plots of Consensus v. Observed Probabilities in the English & Scottish Dataset, Level 1. 49
3.12 Plots of Consensus v. Observed Probabilities in the English & Scottish Dataset, Level 2. 49
3.13 Plots of Consensus v. Observed Probabilities in the English & Scottish Dataset, Level 3. 50

3.14 The sum of the $\mathbb{P}_{\text{cons}}$(1X2 Outcomes) v. $\mathbb{P}_{\text{cons}}$(1X2 Home Win), split by Level. 51
3.15 The sum of the $\mathbb{P}_{\text{cons}}$(1X2 Outcomes) v. $\mathbb{P}_{\text{cons}}$(1X2 Home Win), split by Season. 52
3.16 The sum of the $\mathbb{P}_{\text{cons}}$(UO Outcomes) v. $\mathbb{P}_{\text{cons}}$(Over 2.5 Goals), split by Level. 52
3.17 The sum of the $\mathbb{P}_{\text{cons}}$(UO Outcomes) v. $\mathbb{P}_{\text{cons}}$(Over 2.5 Goals), split by Season. 53
3.18 The sum of the $\mathbb{P}_{\text{cons}}$(AH Outcomes) v. $\mathbb{P}_{\text{cons}}$(AH Home Win), split by Level. 53
3.19 The sum of the $\mathbb{P}_{\text{cons}}$(AH Outcomes) v. $\mathbb{P}_{\text{cons}}$(AH Home Win), split by Season. 54

List of Tables

1.1 The UEFA Country Coefficient for the top six European leagues (UEFA.com, n.d.[a]). 10
1.2 The English and Scottish football league pyramids (English Football League, n.d. Scottish Professional Football League, 2021). . 10
1.3 Columns from `football-data.co.uk`'s datasets used. 13

2.1 IDA calculations for the *elite* dataset. 17
2.2 EDA calculations for the *elite* dataset 21
2.3 Matches with $\mathbb{P}_{\text{cons}}(\text{Draw}) > 0.6$ 23
2.4 The Bin Size for each tile of Figure 2.5 24
2.5 R^2 and RMSE values for the elite leagues, 2005-20 27
2.6 Standard errors for each outcome in our model. 28
2.7 P_1 and P_2 value for all elite leagues. 29
2.8 R^2, RMSE, P_1, P_2 for all *elite* leagues. 29
2.9 Data from 1963/64-2004/05 Seasons (Goossens, 2005). 31
2.10 R^2, RMSE, P_1, P_2 for our *elite* league dataset, split by season. . 32
2.11 By-League PCA values. 34
2.12 By-Season PCA values. 36

3.1 EDA calculations 40
3.2 The Bin Size for each tile of Figures 3.4 and 3.5 44
3.3 Linear models to predict the observed outcome $\mathfrak{O}$ from a given bookmaker consensus probability $\mathfrak{C}$ using the English/Scottish Data. 46
3.4 Values for R^2 and RMSE for the markets of interested, based on the models in Table 3.3 46
3.5 R^2 and RMSE values for the 1X2, UO and AH markets across all three levels, and P_1 and P_2 for the 1X2 market. 48
3.6 The mean overroud $\bar{\eta}$ for different groups of our dataset. 54

0.2 *Disclaimer*

Involvement in gambling can often lead to highly dangerous and damaging consequences, both for the bettor and those around them: problem gambling is characterised as "*persistent and recurrent problematic gambling behaviour leading to clinically significant impairment or distress*" (American Psychiatric Association, 2018).

This dissertation will not seek to solve the social, economic, or political issues surrounding gambling, but will instead focus wholly on the mathematics and statistics upon which sports gambling is based.

Chapter 1

Introduction

1.1 Background Information

1.1.1 Association Football in Europe

Association football—also known as 'soccer' or simply 'football'—is the most popular sport in the world (Giulianotti, 2012), with, according to a survey by the world's governing body, FIFA, in 2001, over 240 million players worldwide (FIFA.com, 2001).

Professional football in Europe is governed by UEFA, with each nation having their own governing body (such as the Deutscher Fußball-bund in Germany) administering football in that country (UEFA.com, n.d.[b]), including the league systems/pyramids. Generally, the systems follow a similar structure, with a number of promotion and relegation places contested for throughout the course of a year-long season, ensuring each division has a similar ability (The Football Association Premier League Limited, 2019). The top division in each country is allocated a number of qualification places to the two Europe-wide club competitions: the Champions' League and Europa League.[1] This allows the best teams from each country to compete against each other.

The number of allocated places is chosen via the UEFA country coefficient, which ranks the countries by the performance of their collective clubs in these competitions: we choose the top six[2] of these to be considered the *elite* European leagues. These are given in Table 1.1.

Later in this dissertation, we will consider the English and Scottish football pyramids; the order of their leagues used in the analysis are given in Table 1.2. We choose these two leagues as `football-data.co.uk` have data on five English leagues and four Scottish leagues. No other league has more than two; this will allow us to compare between different *tiers* in the pyramid.

[1] A third competition, the Europa Conference League, is planned for the 2021/22 season (UEFA.com, 2020).

[2] As of 5th January, 2021.

Table 1.1: The UEFA Country Coefficient for the top six European leagues (UEFA.com, n.d.[a]).

Country	*Top Division*	*Coefficient*
Spain	La Liga	92.283
England	Premier League	90.712
Italy	Serie A	72.295
Germany	Bundesliga	71.856
France	Ligue Une	54.915
Portugal	Premiera Liga	47.349

Table 1.2: The English and Scottish football league pyramids (English Football League, n.d. Scottish Professional Football League, 2021).

Tier in Pyramid	*English Pyramid*	*Scottish Pyramid*
1	Premier League (EPL)	Premier League (SPL)
2	Championship	Division 1
3	League One	Division 2
4	League Two	Division 3
5	Conference/National League	—

1.1.2 Probabilities, Odds & Gambling

Every event, say i, has a PROBABILITY[3] of occurring, denoted as $\mathbb{P}(\text{Event } i) = p_i$, between 0 (almost never occurs) and 1 (almost certain to occur), with the sum of all possible outcomes being equal to 1 (Grinstead and Snell, 2012). For example, when rolling a fair, six-sided die, numbered one to six, we can draw a table with the probabilities of rolling the die and number X being rolled:

x	1	2	3	4	5	6
$\mathbb{P}(X = x) = p_i$	1/6	1/6	1/6	1/6	1/6	1/6

In gambling, the ODDS of an event are used, rather than the probabilities (Štrumbelj, 2014). For a completely fair, non-profit casino or bookmaker, the odds offered for an outcome would be equal to the inverse of the probability of the event: for our fair die scenario above, the odds O_i (written in different styles[4]) would be:

	x	1	2	3	4	5	6
O_x	British Style/Fractional	6/1	6/1	6/1	6/1	6/1	6/1
	European Style/Decimal	7.0	7.0	7.0	7.0	7.0	7.0
	American Style/Moneyline	600	600	600	600	600	600

If one was to place a bet worth £1 (the STAKE) on rolling a three, and this occurred, the casino or bookmaker would—for British and European odds—multiply the stake, S, by the odds O to find one's profit: the total pay-out, P

[3]Throughout this dissertation, several words have been written in SMALL CAPS: these are defined in Appendix A.

[4]Throughout this report, we will use the European/decimal odds system: this is also favoured by `football-data.co.uk`.

would include the original stake: $P = SO + S$. A positive American odd shows the amount of money you would win for a \$100 bet (or whichever currency), whereas a negative odd is the amount you need to bet in order to win \$100 (Cronin, 2019).

1.1.3 Betting Markets of Interest

There are a vast range of MARKETS used in football betting (bet365, n.d.); the ones we are most focused on in this report, are:

- 1X2 (or full-time result)—betting on the final outcome of the match being either a home win, a draw, or an away win.
- GOAL MARKETS—this is a bet on the amount of total goals or number goals for either side, usually given as a half (E.G., 0.5, 1.5, 2.5, ETC.), with odds offered for under/over the given amount.
- ASIAN HANDICAP (A.H.)—this style of betting allows a seemingly one-sided fixture to become competitive. For A.H. bets, a handicap given can be a whole number or half-number, used as a headstart (if positive) or a detriment (negative) to a team; for example, if, in a match between Chelsea and West Brom, Chelsea's handicap was -2.5, they would need to win by 3 goals in order for a bet on them to win: West Brom would just need to avoid a loss by 3 goals for a bet on them to win. In addition, Asian handicaps can also be quarter-numbers (say $\frac{3}{4}$): the bet is then split into two: half the stake on $\frac{1}{2}$ and half on 1, in this case.

Gambling is a huge part of football culture, with 27 of the 44 teams (61.3%) in the English Premier League and Championship having a gambling company as their main shirt sponsor (Davey, 2020); those that don't likely have a betting company as a 'Club Partner', for example Arsenal, who have Fly Emirates on their shirts, are partnered with SportsBet.io; Manchester City—Etihad Airways on their shirts—are partnered with Marathon Bet (The Football Association Premier League Limited, 2019). The combined income of betting partnerships in the Premier League is around £70 million. 8 of the 20 Spanish La Liga sides have a gambling company on their shirts: BetWay alone sponsor three (Score and Change, 2020). This leads one to wonder: how accurate are the betting odds by these companies, so invested in football, on the matches?

1.2 Literature Review

1.3 Rationale, Aims, Objectives and Methods

Throughout this project, we will aim to assess the accuracy of betting odds offered by a range of bookmakers in European football. We will look into different markets: the 1X2, A.H. and goals markets, and try to determine factors that may or may not influence the accuracy of the odds, such as the level of the match and country (league) of the match. We will also look briefly into the OVERROUND: where the odds are lowered, allowing bookmakers to make profit.

In order to answer this, we will utilise data from `football-data.co.uk`—a website set up by Joseph Buchdahl, a betting analyst and author of multiple published works about betting (Buchdahl, n.d.[b])—which provides historical results and odds in easy-to-access comma-separated-value (`.csv`) files (Buchdahl, n.d.[c]). With this data, we will explore different angles assessing the accuracy of the odds, including exploratory analysis (Hoaglin, 1977) and calculating the predictive power of the odds (Owen, 2009). In addition, we will look at a range of visual aids such as tile plots, histograms, and density plots. We will create models to predict the actual observed probability of an event from the bookmaker consensus probability, obtained by taking the inverse of the consensus odds and normalising, that is, ensuring the sum equals one. The equation for normalisation is given in Equation 1.1, where j is the outcome being normalised; the summation is for all i out of k outcomes, E.G., for the 1X2 market, $k = 3$: a home or away win, or a draw; for the Under/Over 2.5 goals market, $k = 2$: under or over.

$$\frac{\mathbb{P}_{\text{cons}}(\text{Outcome}_j)}{\sum_{i=1}^{k} \mathbb{P}_{\text{cons}}(\text{Outcome}_i)} \tag{1.1}$$

.

Throughout the project, we will utilise the programming language R and a range of additional packages (Section 1.3.1) allowing us to perform these tasks efficiently (R Core Team, 2021).

Our objectives are as follows: first, we will look into the accuracy of odds offered in the 1X2 market in *elite* European leagues: the Spanish La Liga, English Premier League, Italian Serie A, German Bundesliga, French Ligue Une, and Portuguese Premiera Liga. We will combine this analysis with a review into whether or not COMPETITIVE BALANCE impacts the accuracy of the odds offered.

Secondly, we will look into the accuracy of odds offered in different levels of the English and Scottish league systems, assessing whether the accuracy of odds differs between the levels (and standards) of football, looking into three markets: the 1X2, A.H. and Under/Over 2.5 goals. We will combine this with an investigation into the OVERROUND, a measure of commission bookmakers take.

Finally, we will develop, and test, a method of placing bets, using our previous findings to assist in the model creation, and testing it against a random bet strategy: placing a similar number of bets at random.

Upon the completion of our objectives, we will conclude with a discussion into our findings and proposing areas for future research.

1.3.1 R Packages Used

`car` — "Companion to Applied Regression" (Fox and Weisberg, 2019).
`MASS` — Support for statistical functions and tests, such as the χ^2 test of independence (Venables and Ripley, 2002).
`ggplot2` — For elegant graphics and a wide range of plots and graphs. (Wickham, 2016).
`gridExtra` — To allow for multiple-figure plots created by `ggplot2` (Auguie,

2017).
`scales` — "Scale functions for visualisation" (Wickham and Seidel, 2020).
`e1071` — Miscellaneous statistical functions, used in this project for the `discrete` distribution tools (Meyer et al., 2020).

1.3.2 The Data

Sourcing the data

For this project, we have used `football-data.co.uk` (F-D) to collect data. F-D use the following bookmakers for their odds (as of 02/10/20): Bet365, Blue Square Bet, Bet & Win, Gamebookers, Interwetten, Ladbrokes, Pinnacle, Sporting Odds, Stan James, BetVictor, William Hill. For game statistics, such as home/away corners, free kicks, shots (on/off target), offsides and cards, F-D uses BBC Sport, ESPN Soccer, Gazzetta.it and Football.fr, with betting odds taken from the individual bookmakers. The odds for matches during the weekend are collected on Friday afternoons; odds for midweek matches are collected on Tuesday afternoons. Statistics for 2000-01 and 2001-02 for the English, Scottish and German leagues were provided by Sports.com (which is under new ownership and now unavailable) (Buchdahl, n.d.[a]).

The columns from the F-D `.csv` files that are required for our analyses are given in Table 1.3.

Table 1.3: Columns from `football-data.co.uk`'s datasets used.

Name	*Meaning*
`div`	Division
`date`	Date of the match
`HomeTeam`, `AwayTeam`	The home/away side in the match
`FTHG`, `FTAG`	Full-time home/away goals
`FTR`	Full-time result: Equal to 'H' for a home win, 'A' for an away win, and 'D' for a draw.
`BbAvH`, `BbAvA`, `BbAvD`	The bookmaker consensus odds for a home win, away win and draw, respectively. From the 2019/20 seasons onwards, these were renamed `AvgH`, `AvgA` and `AvgD`.
`BbAv>2.5`, `Avg>2.5`	(Rendered as `BbAv.2.5` and `BbAv.2.5.1` in R.) The bookmaker consensus odds for Over 2.5 Goals and Under 2.5 Goals. Renamed `Avg>2.5` and `Avg<2.5` from 19/20.
`BbAvAHh`	The bookmaker consensus home handicap offered.
`BbAvAHH`, `BbAvAHA`	The bookmaker consensus odds for the Asian Handicap market.

Data Storage

With R, we can access the `.csv` files without downloading them, using the URL in place of a file name with the `read.csv("")` command. Further, we can use a `for` loop to download and store all the datasets in R; this is shown in Section 2.1.

The country codes given to each country are:

Spain	`es`	France	`fr`
England	`en`	Portugal	`po`
Italy	`it`	Scotland	`sc`
Germany	`de`		

1.4 Structure

Following this introduction, we split the main body of the dissertation into three chapters: Chapters 2, 3 and 4, before finishing with a conclusion (Chapter 5), where we discuss our findings (Section 5.1), before I discuss some challenges I found throughout the project, and how I overcame them in Section 5.2.

In Chapter 2, we assess the accuracy of betting odds in *elite* European leagues (defined in Table 1.1) in the 1X2 betting market. We begin by conducting initial (Section 2.1) and exploratory (Section 2.2) data analysis. These steps will help us assess the suitability of our data, find general trends, and locate areas for further analysis. After this, we conduct correlation analysis (Section 2.9) where we create linear models with a view to predict the observed actual probability of an event, based on the bookmaker consensus probability. In Sections 2.4, 2.5, and 2.7, we use statistics found in our correlation analysis, as well as predictive power statistics P_1 and P_2 (Owen, 2009) to compare the accuracy across leagues and seasons, investigating possible reasons behind this; in particular, we look into competitive balance across the *elite* leagues in Section 2.6, and conduct principal components analysis in Section 2.8. Finally, we conclude the chapter in Section 2.9.

In Chapter 3, we conduct similar analysis. We begin with exploratory data analysis (Section 3.1), where we find basic calculations and use visual analysis to find trends and areas for further study. We then conduct correlation analysis on the data, finding differences and similarity in bookmaker accuracy across different levels. We conclude the chapter by a brief look into the overround, and where bookmakers apply the largest percentage commission.

In Chapter 4, we use our findings from Chapter 2 and 3 to construct a *method* of placing bets, with the aim of making long term gain on an investment. We first outline our method, which is based on winning as many bets as possible, without falling for "attractive" odds. Once the model is defined, we run it on our entire dataset, across our selected markets, and review its performance. We compare it to a *random bet strategy*, which we run multiple times before concluding.

Finally, in Chapter 5, we make our final remarks. We discuss what we have found, how it can be used, and areas for future research in Section 5.1, and I discuss what I personally have learnt throughout the course of the project in Section 5.2.

Appendices

Throughout the dissertation, several words are written in SMALL CAPS: these are defined in A, with mathematical terms in Section A.1 and gambling terms in Section A.2. Located in Appendices B, C, and D is the full R code used throughout the project. The project diary is in Appendix E, the dissertation word count is in Appendix F, and the ethical approval certificate is in Appendix G.

Chapter 2

Assessing the accuracy of betting odds in *elite* European football leagues, from 2005 to 2020.

This chapter will aim to assess the accuracy of betting odds offered by bookmakers in *elite* European leagues, as defined previously (Section 1.1.1, Table 1.1). To begin our analysis—after reading in and cleaning our data—we will use the techniques of initial and exploratory data analysis, with the former looking at a smaller sample of data and ensuring it behaves as expected and the latter looking at four "major themes" (Hoaglin, 1977) to answer questions about our dataset. We will also create visual aids to further explore and analyse the dataset.

Then—using correlation analysis—we will create models to predict the observed probability of a result from the bookmaker's offered odds: from these models we will be able to find statistics to test the FIT of the data, such as the root-mean-squared-error. To aid this, we will find the P_1 and P_2 statistics for predictive performance (Owen, 2009), before looking at the impact the COMPETITIVE BALANCE of a league has on the accuracy of the bookmaker and presenting our conclusions. Finally, we conduct principal components analysis, as a further method of analysing the impact the league and season.

2.1 Initial Data Analysis

Initial data analysis (IDA) is used to answer four questions about our data:

- What is the quality of our data?;
- What is the quality of the measurements?;
- Did the implementation of the study fulfil the intentions of the research design?;
- What are the characteristics of the data sample? (Adèr, 2008)

To answer these questions, we will conduct simple tasks—such as calculating the mean consensus probabilities and comparing them with the observed probabilities, and plotting a histogram to assess the distribution of the consensus probabilities—on a sample data set: we choose one league for one season. Chosen at random, we will look into the **French Ligue Une** over the **2016/17** season.

Our first step is reading in the data from `football-data.co.uk`, and removing the columns we do not need. We use the code below to do this.

```
fr_l1_1617 <-
  read.csv("https://www.football-data.co.uk/mmz4281/1617/F1.csv")
fr_l1_1617 <-
  fr_l1_1617[,c("Div", "Date", "HomeTeam", "AwayTeam", "FTHG", "FTAG",
                "FTR", "BbAvH", "BbAvD", "BbAvA")]
fr_l1_1617 <- na.omit(fr_l1_1617)
```

As mentioned in Section 1.3, we will be considering probabilities throughout the analyses, rather than odds: our first step is to calculate the normalised (see Section 1.3, Equation 1.1) bookmaker consensus probabilities; to perform this using R, we use the code below.

```
fr_l1_1617$AvgHProbPN <- with(fr_l1_1617, round(1/BbAvH, 4))
fr_l1_1617$AvgDProbPN <- with(fr_l1_1617, round(1/BbAvD, 4))
fr_l1_1617$AvgAProbPN <- with(fr_l1_1617, round(1/BbAvA, 4))
fr_l1_1617$Overround <- with(fr_l1_1617,
                             (AvgHProbPN + AvgDProbPN + AvgAProbPN))
fr_l1_1617$AvgHProb <- with(fr_l1_1617, round(AvgHProbPN/Overround,4))
fr_l1_1617$AvgDProb <- with(fr_l1_1617, round(AvgDProbPN/Overround,4))
fr_l1_1617$AvgAProb <- with(fr_l1_1617, round(AvgAProbPN/Overround,4))
```

We normalise the probabilities to counter the bookmaker commission, or *overround*, despite this differing between games and between outcomes (Henery, 1999). This allows us to treat the consensus probabilities as statistical probabilities.

To begin our analysis, we will compute the consensus mean probabilities μ_i and consensus standard deviation σ_i for each outcome i, in order to compare with the observed probabilities, and to compare variation in each outcome. The results, to four decimal places, are shown in Table 2.1.

Table 2.1: IDA calculations for the *elite* dataset.

	Home Win	*Draw*	*Away Win*
Mean Consensus Probability μ_i	0.4414	0.2701	0.2886
Observed Probability	0.4895	0.2474	0.2632
Consensus Standard Deviation σ_i	0.1540	0.0469	0.1366

From this table, we can see the mean consensus probabilities are very close to the observed probabilities: less than 0.05 off for each outcome. Interestingly, the standard deviation of a draw is much lower than the standard deviation for both home and away wins, indicating that the variation in the odds offered for

draws are much lower than those for a clear winner[1].

To visualise these distributions, we produce histograms of the consensus probabilities, shown in Figure 2.1. As a demonstration, we create the histogram for the consensus probabilities of a home win, $\mathbb{P}_{\text{cons}}(\text{Home Win})$, using the code below.

```
ggplot(fr_l1_1617, aes(AvgHProb)) +
  geom_histogram(binwidth=0.05, fill="blue") +
  coord_cartesian(xlim=c(0,1)) + theme_light() +
  labs(title="Home Win", x=NULL, y=NULL)
```

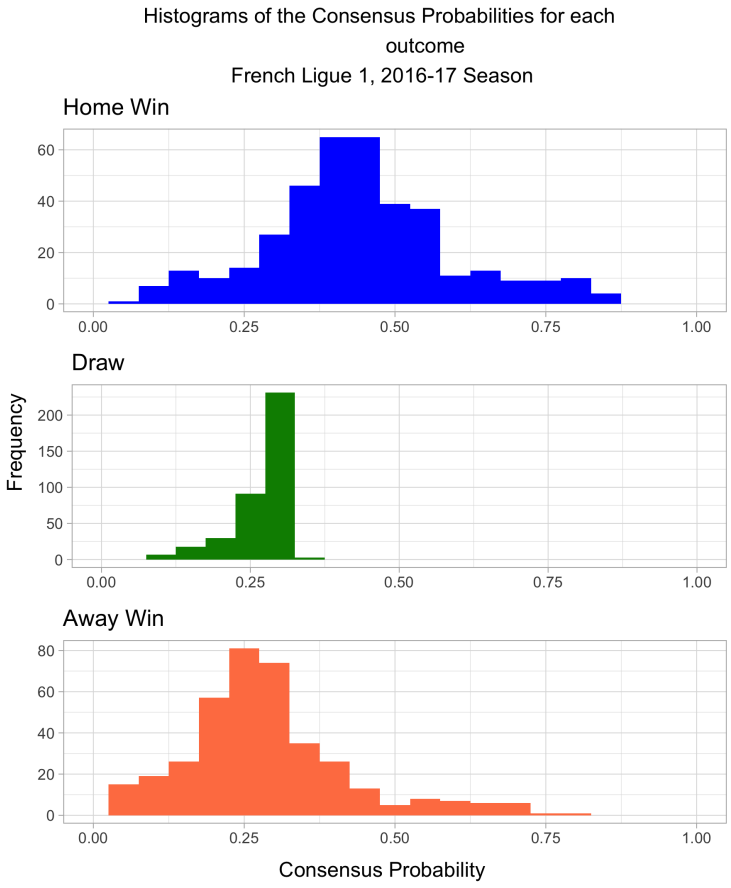

Figure 2.1: Histograms of the consensus probabilities for each outcome, in the French Ligue Une for the 2016/17 season. N.B. the change of scale on the y axes.

This figure shows that the consensus probabilities of a home win are symmetrically distributed around the mean $\mu_{\text{H}} = 0.44$, with a distribution similar to the Normal bell curve. For the consensus probabilities of an away win, the

[1] I.E. a home or away win.

probabilities are positively skewed[2] suggesting a greater proportion of measurements lie to the right/are greater than the peak value (Mendenhall, Beaver, and Beaver, 2013). This means our data will include a few unusually large measurements. Contextually, these could be the league leaders playing away against a relegation-battling side.

The $\mathbb{P}_{\text{cons}}(\text{Draw})$ graph display assists our finding (based on σ_{Draw}) that the variation of bookmaker consensus probabilities of a draw is very low: Over 225 ($N = 380$ matches) of the games lie in the same bin[3] with no values at all recorded above 0.4.

The rationale behind conducting IDA was to answer the four questions set out above: about the quality of our data and measurements, whether the implementation of the study fulfilled the intention of the research design, and the characteristics of our sample (Adèr, 2008). The quality of the data and measurements are both acceptable for our research: no measurements are missing, our values were as expected ($\mathbb{P}_{\text{cons}} \approx \mathbb{P}_{\text{obs}}$), and we were able to conduct computations and produce plots without any issues. The question of the study, assessing the assessing the accuracy of betting odds, will be able to be answered: we have access to the information we need (consensus odds and actual results), in fact, we have much more information we can read in if needed. And our final question, in reference to the characteristics of our sample, can be answered by saying that our data suggests the mean consensus probabilities are roughly equal to the observed probability for each outcome; bookmakers offer less variation on their odds for a draw than for a clear winner; and the consensus probability of a home win is symmetric, whereas the consensus probability of an away win is positively skewed.

2.2 Exploratory Data Analysis

In this section, we initially explore the data. There are four "major themes which motivate many of the techniques" we will apply:

- **Displays**: reveal major features, outliers, non-linearities, discontinuities, skewness, ETC. that calculations such as means, standard deviations and least square regressions cannot show;
- **Residuals**: defined as the observed data minus the fitted data, a clear pattern in the plot of the residuals v. fitted values indicates improvement is possible;
- **Resistance**: dealing with outliers. If we find them, we can run parallel tests (one with outliers; one without) and comparing, testing the resistance of our data to the outliers (not dissimilar to the idea of statistical leverage);
- **Transformations**: does adding a transformation, such as taking the n-th root, logarithms, logits/probits, ETC. allow us to make sense of the data. (Hoaglin, 1977).

[2] Or right-skewed.

[3] The width of all bins across all three histograms is 0.05.

In this section, we will also create visual analyses of the bookmaker's odds and performance; namely, boxplots, density plots, and a tile plot. To begin, however, we must first read in (as we will be considering six leagues and 15 seasons (90 datasets) it is efficient to use a `for` loop to do this) and clean our data. The code to do so—and to normalise our underlying probabilities—is below.

```
#Define which countries and seasons we need to read:
countries <- c("de", "en", "es", "fr", "it", "po")
co.we <- c("D1", "E0", "SP1", "F1", "I1", "P1")
#n.b. The Premier League's code is 0; other countries are 1.
seasons <- c("0506", "0607", "0708", "0809", "0910", "1011", "1112",
             "1213", "1314", "1415", "1516", "1617", "1718", "1819",
             "1920")
eliteTemp <- NULL; elite <- NULL
for (i in seasons){
   for (j in 1:6){
     eliteTemp <- read.csv(paste0(
       'https://www.football-data.co.uk/mmz4281/',i,'/',co.we[j],'.csv'),
       fileEncoding = 'latin1')
     eliteTemp$Country <- with(eliteTemp, countries[j])
     eliteTemp$Season <- with(eliteTemp, i)
     if (i=="1920"){
       eliteTemp$BbAvH<-eliteTemp$AvgH; eliteTemp$BbAvA<-eliteTemp$AvgA
       eliteTemp$BbAvD<-eliteTemp$AvgD
     }
     else{}
     eliteTemp <- eliteTemp[ ,c("Div", "Date", "HomeTeam", "AwayTeam",
                                "FTHG", "FTAG", "FTR", "BbAvH", "BbAvD",
                                "BbAvA", "Country", "Season")]
     elite <- rbind(elite, eliteTemp)
   }
}
elite <- na.omit(elite)

#Finding underlying probabilities:
#Pre-Normalised Probabilities
elite$AvgHProbPN <- with(elite, round(1/BbAvH, 4))
elite$AvgDProbPN <- with(elite, round(1/BbAvD, 4))
elite$AvgAProbPN <- with(elite, round(1/BbAvA, 4))
#To normalise them:
elite$overround<-with(elite, (AvgHProbPN + AvgDProbPN + AvgAProbPN))
elite$AvgHProb <-with(elite, round(AvgHProbPN/overround, 4))
elite$AvgDProb <-with(elite, round(AvgDProbPN/overround, 4))
elite$AvgAProb <-with(elite, round(AvgAProbPN/overround, 4))
```

The number of matches in our `elite` dataset, $N = 31,346$. As this is a large set, we can apply the CENTRAL LIMIT THEOREM (CLT) and assume the mean of the random variables (contextually, matches) follows the NORMAL DISTRIBUTION. For our later analysis, we also will need to find the *correct* probability (and the natural logarithm of it) (the bookmaker consensus probability of the event that was observed) and the two *incorrect* probabilities.

Our first step after cleaning our data is to compute, as in the IDA, the

bookmaker consensus mean probabilities and compare them to the observed probabilities. These are shown in Table 2.2.

Table 2.2: EDA calculations for the *elite* dataset

	Home Win	*Draw*	*Away Win*
Mean Consensus Probability μ_i	0.4472	0.2620	0.2908
Observed Probability	0.4589	0.2566	0.2845
Consensus Standard Deviation σ_i	0.1714	0.0478	0.1536

As seen in Table 2.1, the mean consensus probabilities are extremely close to the observed probabilities: the magnitude difference is than 0.02 for all three outcomes. Similarly, the standard deviations indicate that the consensus probabilities offered for a draw vary significantly less than for those with a clear winner.

Our first graphical step with the elite dataset is to create a boxplot for each outcome. This will allow us to see whether or not there is a significant difference in the observed outcome for the bookmaker probabilities of each event, as well as possibly seeing a trend in the bookmaker's consensus probabilities. This is shown in Figure 2.2.

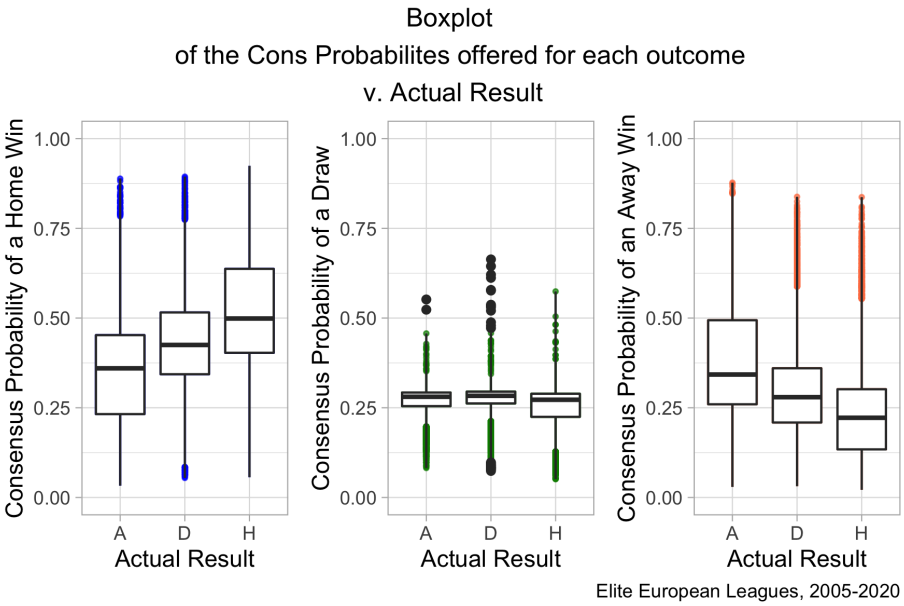

Figure 2.2: Boxplots of the consensus probabilities offered for each outcome.

The figure implies there is no significant difference: whilst it appears there is a correlation—for games that end in home wins, the consensus probability of a home win is higher, and vice versa—this plot shows no strong significance. However, it does strongly reiterate our previous point about variation (or lack

thereof) of the consensus probabilities of a draw (the boxes are much smaller). Interestingly, however, the draw plot has a large number of outliers. We will consider these later.

Instead of creating a binned histogram, as in Figure 2.1, we will use a density plot, which allows for a nicer way of visualising the data, and will allow us to easily compare between leagues: using KERNEL DENSITY ESTIMATION, we will smooth out a histogram using a number of equally spaced points at which the *density* of the distribution is estimated. This number of spaces is a power of two: we choose $2^9 = 512$ (and a Normal (Gaussian) kernel, due to the CLT). This is achieved using the code below: this shows our plot for home and away wins: we include draws separately due to a different y axis scale being used. This results in Figure 2.3.

```
ggplot(elite, aes(x=AvgHProb)) +
  geom_density(color="blue") +
  geom_density(data=elite, mapping=aes(x=AvgAProb), color="coral",
               show.legend=T) + coord_cartesian(xlim=c(0,1)) +
  labs(title="Home and Away Wins", caption="Elite Leagues, 2005-2020",
       x="Consensus Probability", y="density") + theme_light()
```

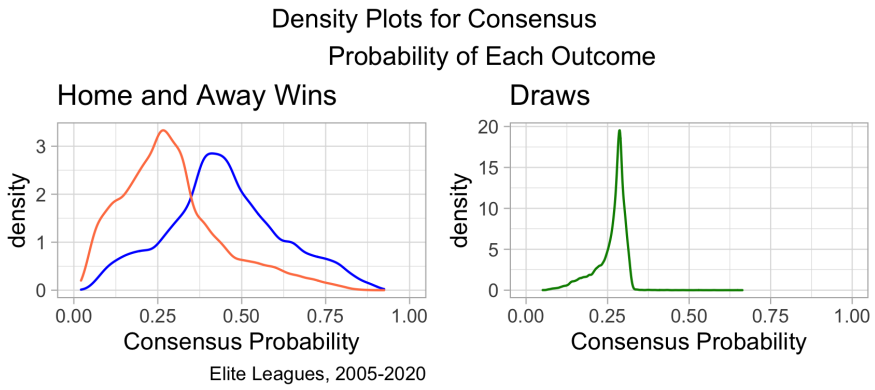

Figure 2.3: Density plots for consensus probabilities offered for each outcome in the 1X2 market.

Interpreting this figure, we can immediately see that the trend shown from Figure 2.1—$\mathbb{P}_{\text{cons}}$(Home Win) is symmetrically distributed around the peak ($\mu_H = 0.45$); $\mathbb{P}_{\text{cons}}$(Away Win) has a positive skew; and $\mathbb{P}_{\text{cons}}$(Draw) has much less variation, with few matches above the peak—still holds. However, we observe a number of matches with a large consensus probability of a draw (up to $\mathbb{P}_{\text{cons}}$(Draw)$\approx$0.65). Table 2.3 shows matches with a consensus probability of a draw greater than 0.6[4].

Immediately, one notices all of these matches are both a) in the Italian Serie A; and b) in the late stages of the football season[5], leading us to two possible

[4] 0.6 is seven standard deviations away from the mean consensus draw probability.
[5] The season normally starts in August and ends in May (The Football Association Premier

Table 2.3: Matches with $\mathbb{P}_{\text{cons}}(\text{Draw}) > 0.6$

League	*Date*	*Home Team*	*Away Team*	*FTR*	$\mathbb{P}_{\text{cons}}$(Draw)
Serie A	09/05/10	Bologna	Catania	1-1	0.6634
Serie A	08/05/11	Bologna	Parma	0-0	0.6445
Serie A	20/05/07	Torino	Livorno	0-0	0.6208
Serie A	03/04/11	Chievo	Sampdoria	0-0	0.6121

reasons behind this:

- The Italian Serie A has a history of match fixing in recent times: the CALCIOPOLI which occurred during the 2004/05 and 2005/06 season involved Juventus, A.C. Milan and Lazio, among others—three of Italy's largest clubs; in 2015, Catania's president was one of several arrested for match-fixing in Serie B matches (Hafez, 2019).

- Due to these games occurring in the late stages, it is possible for a scenario where both teams would benefit from a certain result.[6]

Whilst fixed matches would naturally impact our results, due to the small number of games impacted, it is unnecessary to exclude them from our analysis, in fact: they could improve our discussion.

As these games are all in the Serie A (in fact, of the 81 matches in our dataset with a consensus probability of a draw greater than 0.35, 75 were Italian matches), it makes sense to split this density plot into the different leagues. The Home Win (and to a lesser extent Away Win) plots in Figure 2.4 imply leagues can be split into two categories: those with a *unimodal* density (one peak), and those with a *trimodal* density (three peaks). The latter group (England, Portugal, Spain) have a peak between 0 and 0.25 (very low probability of a home/away win, depending on the plot) and a peak between 0.75 and 1 (very high probability of a home/away win), whereas the former (Germany, France, Italy) only has the central peak. One reason could be COMPETITIVE BALANCE: we investigate this further in Section 2.6.

The final visual aid in this section we will use is a *tile plot*: similar to a heat map, this will allow us to create a three-dimensional representation of data in the 2D plane. For our dataset, we will plot the match result on the x (home goals) and y (away goals) axes, allowing us to see both the full time result (tiles on the diagonal $x = y$, are draws; upper triangle are away wins $y > x$; lower triangle are home wins $x > y$) and the *magnitude* of the result, or how *convincing* the result is: a match further from the diagonal has a greater disparity in goals, and so can be considered a more convincing win. Representing the z axis, each square will be shaded in with the mean correct bookmaker probability for that result: with a low consensus probability, the square will be lighter. We would expect these to be closer to the diagonal. Due to a low number of *extreme* results (more than 5 goals scored for a team), we will group these into a 5+

League Limited, 2019).

[6]The The Disgrace of Gijón in the 1982 World Cup is a particularly famous example of this (See Appendix A.2, Definition 5).

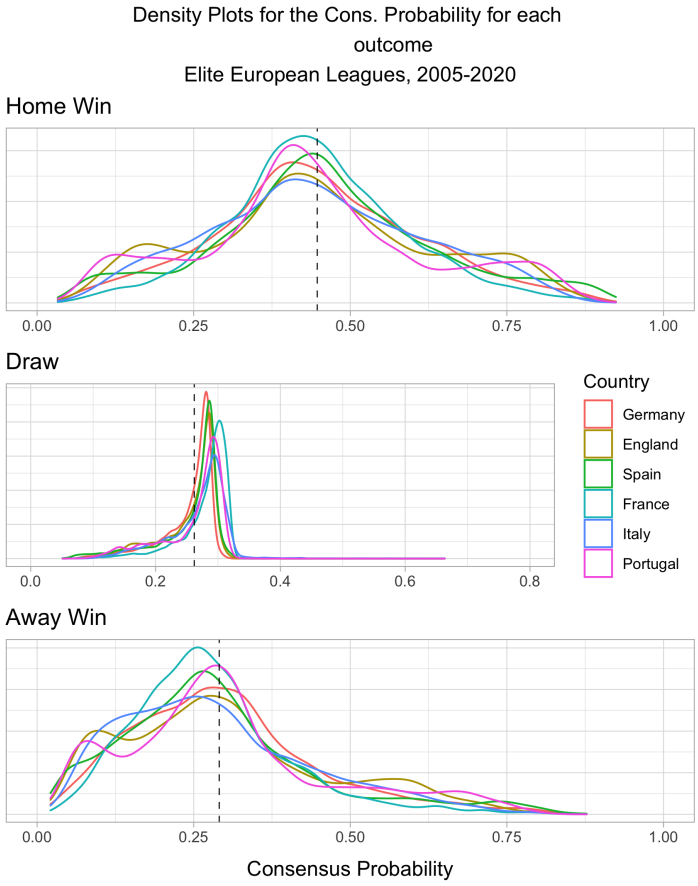

Figure 2.4: Density plots of the consensus probabilities offered for each outcome of the 1X2 market, split by league.

tile[7]. Before we can analyse the plot, it is important to know the bin sizes for each tile. These are given in Table 2.4.

Table 2.4: The Bin Size for each tile of Figure 2.5

Away Goals							
	5+	119	105	67	26	7	2
	4	257	294	174	63	36	12
	3	636	869	547	299	109	45
	2	1451	2003	1508	712	265	106
	1	2299	3688	2756	1375	528	281
	0	2510	3342	2575	1371	594	315
		0	1	2	3	4	5+
		Home Goals					

[7]Our highest scoring draws were 5-5, occuring twice: Lyon v. Marseille (2009) and West Bromich Albion v. Manchester United (2013).

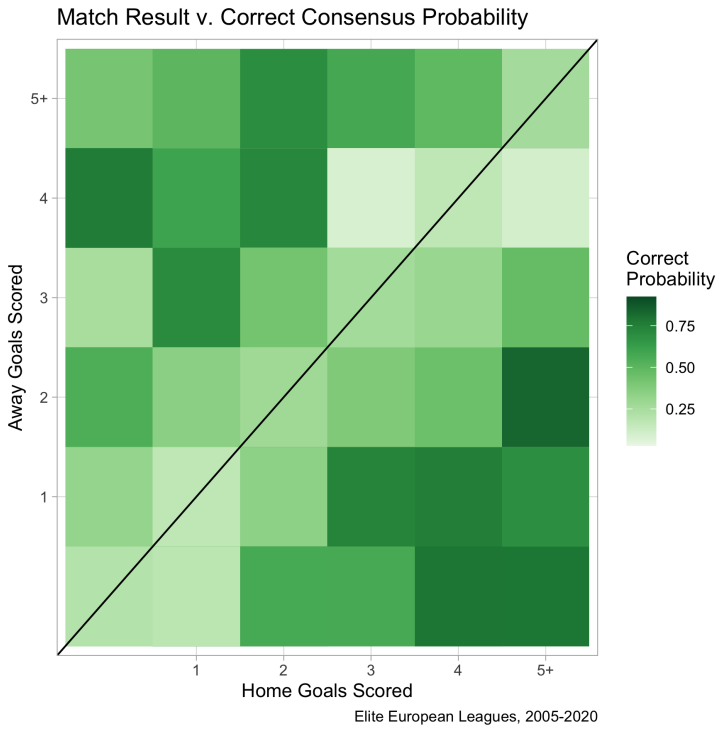

Figure 2.5: Tile plot of the correct consensus probability for each possible result.

Inspecting the figure, we notice that the results on the diagonal all have a similar consensus probability, around 0.3, as one would expect. Considering the lower triangle (home wins) first, we notice the pattern expected holds: that is, tiles furthest from the diagonal are darker. Interestingly, the darkest tile is for a 5 (plus) - 2 home win; the bin size for this is low, though, with $N = 106$ matches.[8] Whilst there are only 12 games in the 5 (plus)-4 tile, it is striking that it has one of the lowest correct consensus probabilities of all the tiles. Finally, considering the upper triangle (away wins), our pattern is not as consistent as with home wins, but there is still evidence of it, with the highest probabilities being further from the diagonal. As a whole, the figure is implying what we would expect: games with greater disparity in full time goals (i.e., more convincing wins) have a higher bookmaker consensus probability of the correct result than closer games and draws.

We conducted EDA to check four themes: displays, residuals, resistance and transformations.

- Displays: Whilst there was no discontinuity, we saw—as in our IDA in Section 2.1—the consensus probabilities of away wins are positively/right-skewed, and of draws have a very low variance. We saw, and assessed, the outlying matches with a large $\mathbb{P}_{\text{cons}}(\text{Draw})$.

[8] 0.0034% of the total dataset.

- Residuals & Resistance: As we have not created a model (we do this in Section 2.3), we cannot plot the residuals.
- Transformations: Our data, so far, makes sense, and is as expected. There is nothing in the plots to suggest we need, or to justify the use of, any transformations.

2.3 Correlation Analysis, and Model Creation

In this section, we will determine the coefficient of determination R^2 and the root mean square error RMSE. We can do this by creating a linear model to predict the observed probability, from a given consensus probability. For high levels of accuracy, we will find a high R^2 and low RMSE (Mendenhall, Beaver, and Beaver, 2013). There are alternatives to RMSE, such as mean square error (MSE), mean absolute error and median absolute error: RMSE is more sensitive to outliers than mean and median absolute error; we choose to use RMSE due to a relatively low amount of outliers, and its greater theoretical relevance (Hyndman and Koehler, 2006). R^2 describes the percentage of the variation in a variable—in our case, observed probability—due to a predictor: the bookmaker consensus probability (Draper and Smith, 1998).

Once we have determined R^2 and RMSE (both overall and for each outcome), we will determine them for each league, which will be used in Section 2.6. We conduct this by *binning* matches: binning is where we partition the data into groups (the bins), with a sufficient amount to "capture the major features in the data while ignoring fine details" (Knuth, 2006). We will group games with similar consensus probabilities, find the $\mathbb{P}$(Outcome) of the bin (the observed probability) and use the consensus v. observed probabilities in our future analysis (plots and model creations). In R, we create bins using the `cut` command, and the `tapply` command to find the mean value in each bin[9]. Our R code and comments for the Home Wins is below: we choose use 124 bins, meaning that each bin has over 250 matches. Doing this for each outcome and overall, we find our R^2 and RMSE values, shown in Table 2.5.

```
elite$AvgHProb.cut <- cut(elite$AvgHProb, 124, include.lowest=T)
#First, we cut the data into 100 'bins'
levels(elite$AvgHProb.cut) <-
    tapply(elite$AvgHProb, elite$AvgHProb.cut, mean)
#Tapply finds the mean of the bin, rather than taking the midpoint
elite.observed.probabilites.TabH <-
    prop.table(table(elite$FTR, elite$AvgHProb.cut), 2)[c(1, 2, 3),]
#The c(1,2,3) will remove any extra (blank) rows
elite.observed.probabilites.H <- elite.observed.probabilites.TabH[3,]
#[n,]; if n = : 1 Away; 2 Draw; 3 Home (alphabetic)
elite.bookmaker.probabilites.H <-
    as.numeric(names(elite.observed.probabilites.H))
```

These values show that, for the home and away win models, a very large ($> 95\%$) amount of the variation in the model generated is explained by the

[9]The default is the midpoint (RDocumentation, n.d.).

Table 2.5: R^2 and RMSE values for the elite leagues, 2005-20

	Home Win	*Draw*	*Away Win*	*Overall*
R^2	0.98665	0.52008	0.96411	0.83832
RMSE	0.03241	0.20767	0.05166	0.11768

consensus probabilities: this is extremely strong.[10] In addition, the RMSE values for the two are very low, around 0.05 for away wins, and 0.03 for home wins. For draws, however, $R^2 = 0.52$ and RMSE $= 0.21$, indicating that the bookmaker's odds are not as accurate for draws than for clear results; this explains why our overall R^2 and RMSE values are lower than home and away wins, at 83.8% and 0.12 respectively.

To further demonstrate this point, in Figure 2.6, we have a scatter plot of the observed probability v. bookmaker consensus probability using the bins created, along with the linear models (and a 95% CONFIDENCE INTERVAL (CI) around them in light grey), and the $x = y$ line for reference (dashed black).

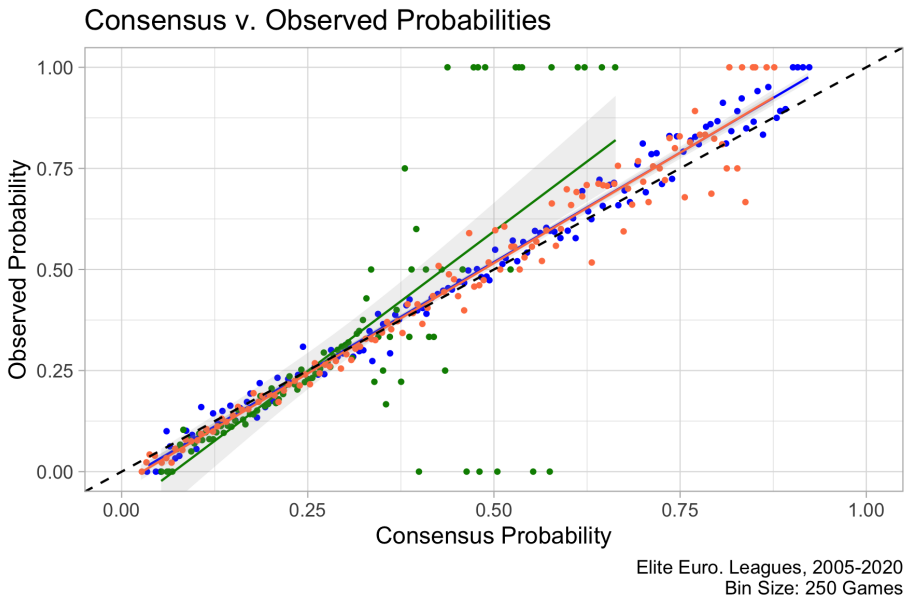

Figure 2.6: Scatter plot of the linear models created.

The figure affirms our observations from the values in Table 2.5, and suggests that the bookmakers accuracy for home and away wins is extremely high. Both the 95% CIs are extremely small, with no major outliers. Noticeable, however, is a group of away win bins with a high consensus probability, but a low observed probability: contextually, this could be down to home advantage, with a teams performance being poor away from home: this is an area for future research. As aforementioned, the bookmaker performance is poor for draws: this is reiterated

[10] For a perfect correlation, R^2=1 (Mendenhall, Beaver, and Beaver, 2013).

in the plot, with a large 95% CI, discussed later. It is worth noting that the accuracy is higher for $\mathbb{P}_{\text{cons}}(\text{Draw}) \in (0, 0.3)$ and remarkably poor for consensus probabilities above this mark: this coincides with the straight drop in density in consensus draw probabilities in Figure 2.3. The values on the $y = 1$ line—where all games in the bin were draws—are likely to be due to the aforementioned match fixing (Section 2.2).

Confidence intervals are calculated by Equation 2.1 (Mendenhall, Beaver, and Beaver, 2013), where $\bar{x}$ is the point estimate, μ is the true value and $Z_{\alpha=0.025}$ is the t statistic as N approaches infinity (using the CLT, we can assume this with a large n). The large CI for the draw model indicates the STANDARD ERROR (SE) for the draw model is high; the SE values for each model are given in Table 2.6.

$$\mathbb{P}\Big(\mu \in \big(\bar{x} \pm Z_{\alpha=0.025} \cdot SE\big)\Big) = 0.95 \tag{2.1}$$

Table 2.6: Standard errors for each outcome in our model.

	Model			
	Home Wins	*Draws*	*Away Wins*	*Overall*
SE	0.011404	0.13835	0.019151	0.02651

The models created in this section, which can be used to predict the observed probability $\mathfrak{O}$ from a consensus probability $\mathfrak{C}$, are in in Equations 2.2 to 2.5.

$$\mathfrak{O}_{\text{Home Win}} = -0.02323 + 1.08278 \cdot \mathfrak{C}_{\text{Home Win}} \tag{2.2}$$
$$\mathfrak{O}_{\text{Draw}} = -0.09646 + 1.38142 \cdot \mathfrak{C}_{\text{Draw}} \tag{2.3}$$
$$\mathfrak{O}_{\text{Away Win}} = -0.03018 + 1.09182 \cdot \mathfrak{C}_{\text{Away Win}} \tag{2.4}$$
$$\mathfrak{O}_{\text{Overall}} = -0.03194 + 1.11148 \cdot \mathfrak{C}_{\text{Overall}} \tag{2.5}$$

2.4 Measuring Predictive Performance

We have two measures for calculating short-term predictive performance, P_1 and P_2, defined in Equations 2.6 and 2.7 respectively (Owen, 2009), with N being the number of matches in the sample; k being the match number; $\mathbb{P}(O_k)$ being the correct/observed bookmaker probability for match k; and $\mathbb{P}(I_{1k})$, $\mathbb{P}(I_{2k})$ being the two incorrect/not observed bookmaker probabilities for match k: for example, if match m finished in a home win, $\mathbb{P}(O_m) = \mathbb{P}_{\text{cons}}(\text{Home Win})$.

$$P_1 = \exp\left\{\frac{1}{N}\sum_{k=1}^{N}\log_e\big[\mathbb{P}(O_k)\big]\right\}, \quad k \in [1,\ N] \tag{2.6}$$

$$P_2 = \frac{1}{N}\sum_{k=1}^{N}\left\{\big[1 - \mathbb{P}(O_k)\big]^2 + \mathbb{P}(I_{1k})^2 + \mathbb{P}(I_{2k})^2\right\}, \quad k \in [1,\ N] \tag{2.7}$$

For better predictive performance, we seek higher P_1 and lower P_2 values, though on their own will not tell us much. We will instead use the values to compare between leagues. These values are shown in Table 2.7.

Table 2.7: P_1 and P_2 value for all elite leagues.

P_1	0.3788607
P_2	0.5776072

2.5 Comparing leagues

In this section, we will compare the six leagues of our *elite* group, before we investigate a possible reason for any disparity in bookmaker accuracy between leagues—competitive balance—in Section 2.6. In order to compare the six elite leagues, we will conduct a near-identical analysis as in Sections 2.3 and 2.4, to find each league's R^2, RMSE, P_1, and P_2 value. The only difference in the correlation analysis is that we choose to add a weight to our bins: we bin the more varied home and away wins into 20 bins (around 200 games per bin) and the less varied draws into 5 bins (800). In R, we will use a `for` loop to create a temporary linear model, extract the R^2 and RMSE values, and create a table with each league'

s values. The results (and their respective ranking, R^2 and P_1 descending; RMSE and P_2 ascending, with higher ranking indicating better performance) are shown in Table 2.8.

Table 2.8: R^2, RMSE, P_1, P_2 for all *elite* leagues.

	R^2	RMSE	P_1	P_2	Average Rank
Germany	0.95489 6th	0.05472 6th	0.36958 5th	0.59429 5th	5.50
England	0.96773 5th	0.04783 5th	0.38454 2nd	0.56661 2nd	3.50
Spain	0.96950 4th	0.04711 4th	0.38267 4th	0.57029 3rd	3.75
France	0.97938 3rd	0.03845 3rd	0.36591 6th	0.60248 6th	4.50
Italy	0.98599 1st	0.03297 1st	0.38287 3rd	0.57085 4th	2.25
Portugal	0.98216 2nd	0.03642 2nd	0.38894 1st	0.55963 1st	1.50

Looking at the average ranking, we can see that the bookmakers performed most accurately in the Portuguese and Italian leagues, followed by the English and Spanish leagues, and least accurately in the French and German leagues. We investigate whether this is due to competitive balance in Section 2.6.

2.6 The effect of competitive balance on bookmaker accuracy

2.6.1 What is competitive balance?

Competitive balance is a concept that weighs heavily on economics. It is defined, by the Cambridge Dictionary, as "the situation in which no one business of a group of competing businesses has an unfair advantage over the others," with a monopoly being a situation with no competitive balance (Cambridge Dictionary, n.d.).

There is a weight of research into competitive balance in American sports, with all of the 'big four' leagues[11] and the MLS (Major League Soccer: the top-tier football league in the USA) having some form of salary cap to promote balance (Slowinski, 2012; NBA.com, 2020; Rosen, 2020; Barrabi, 2020; Goal, 2019). The nature of competitive balance in baseball that "competitors must be of approximately equal 'size' if any are to be successful; this seems to be a unique attribute of professional competitive sports" (Rottenberg, 1956). However, if we look into *elite* European football, this doesn't hold: in the 2015/16 season, for the English Premier League, Manchester United—who finished fifth—had a turnover of £515m; paid £232m in wages, competed against Watford—finished thirteenth—who had a turnover of £94m; wage bill of £58m (Conn, 2017). In 2020, Statista published a ranking of the German Bundesliga teams by market (transfer) value. FC Bayern Munich had a value of €875m; DSC Arminia Bielefeld had a value of €47m (Lange, 2020). This shows that European football may not, in terms of competitive balance, follow the same guides that American sport does.

2.6.2 Quantifying Competitive Balance

We introduce the NAMSI (NAtional Measure of Seasonal Imbalance), shown in equation 2.8. This is, for a league of n teams, the ratio between two standard deviations: σ_{Season} the observed standard deviation of winning percentage for team i, W_i; and $\sigma_{\text{Certainty}}$ the theoretical standard deviation of a certain season, where the team in 1st place wins every game, the team in 2nd wins every game except those against the team in first, ETC., and the team in ith place loses to all teams above them in the table, and defeats all those below them. The NAMSI statistic lies in the range between 0 and 1, with a lower value corresponding to lower seasonal imbalance (Goossens, 2005).

$$\text{NAMSI} = \frac{\sigma_{\text{Season}}}{\sigma_{\text{Certainty}}} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(W_i - 0.5)^2}{n}}}{\sqrt{\frac{\sum_{i=1}^{n}(C_i - 0.5)^2}{n}}} \tag{2.8}$$

NAMSI is a "static measure since it only looks at one season independently of other seasons." In football leagues, a poorly competitively balanced league system would have the same few teams competing for the same places each

[11] The MLB (baseball), NBA (basketball), NFL (American football), NHL (ice hockey)

season. A fluid measure we introduce is the Top K ranking: we let $K = 3$.[12] Our statistic is the number of teams entering the top three in three consecutive years, with data from 1963/64 to 2004/05, with a value of 3 showing perfect imbalance, and 9 showing perfect balance, we call this value κ (Goossens, 2005).

A further method of quantifying imbalance is the Gini coefficient, G, which is often used to compare wealth inequality between different nations (The World Bank, 2018). It measures the ratio of the area between the Lorenz Curve of the country and the $y = x$ (perfect equality) line. The Lorenz Curve is a graphical measure showing the overall income distribution: on the x axis is the cumulated percent of the population, from poorest to richest; on the y axis is the percent of the total wealth of the country held by this x% (Lorenz, 1905).

In Table 2.9, we have the NAMSI, κ and Gini coefficients (Goossens, 2005) with the leagues in our analysis ranked (NAMSI, Gini descending; κ ascending: higher ranking indicates less balance in the league), and other top-tier European leagues included for reference.

Table 2.9: Data from 1963/64-2004/05 Seasons (Goossens, 2005).

	Values					
	NAMSI		κ		Gini	
Germany	0.374	3rd	5.71	4th	0.723	6th
England	0.372	4th	5.79	5th	0.826	3rd
Spain	0.364	5th	5.07	2nd	0.861	2nd
France	0.342	6th	6.00	6th	0.784	4th
Italy	0.418	2nd	5.36	3rd	0.737	5th
Portugal	0.505	1st	4.07	1st	0.898	1st
Greece	0.488	—	4.14	—	0.870	—
The Netherlands	0.494	—	4.36	—	0.888	—
Sweden	0.410	—	6.07	—	0.692	—
Belgium	0.452	—	5.07	—	0.801	—
Denmark	0.412	—	6.43	—	0.581	—

Table 2.9 implies the Portuguese is the least competitively balanced by all measures; the Italian league has a high NAMSI and low κ value, indicating low competitive balance, but—contrastingly—a low Gini coefficient. By the NAMSI and κ values, the French Ligue Une was the most competitive league, and by the Gini coefficient, the German Bundesliga was. Comparing these values to Table 2.8, we see the leagues with the greatest bookmaker performance (Portugal and Italy) have the worst competitive balance, and vice versa: Germany and France have the lowest bookmaker performance, and the highest competitive balance. These findings indicate that there is a possible link between bookmaker accuracy and competitive balance.

If we consider Figure 2.4—which implies the Spanish, Portuguese and English leagues follow a trimodal distribution; the French, German and Italian

[12] "Because in most European countries it are two or three teams that in general are considered to be dominant. Taking up more teams underrates the dominance since the top 4 and 5 often change," (Goossens, 2005).

leagues follow a unimodal distribution—we can infer a link between competitive balance and the distribution. Portugal's unbalanced Primiera Liga has a trimodal distribution, and higher levels of bookmaker accuracy; both Spain's La Liga and England's Premier League also follow the trimodal distribution, and had the median levels of balance and accuracy. The well-balanced German Bundesliga and French Ligue Une both have a unimodal distribution and lower levels of bookmaker accuracy. The only unexpected result here is the Italian Serie A, with low balance and high accuracy but a unimodal distribution.

A study into the Italian and Spanish football, analysing styles of play (offensive or defensive) showed that whilst the Italian league requires strong defensive efficiency to achieve a high ranking; the contrary is true in Spain, where the "best-rewarded strategy consists in improving offensive efficiency" (Boscá et al., 2009). An area for future research lies here: perhaps, the two distributions are due to the styles of play in those countries.

2.7 Comparing Seasons

In addition to running the by-league analysis, we can run a by-season analysis, to investigate whether or not any changes have taken place over time. Our values are in Table 2.10, and plotted in Figure 2.7.

Table 2.10: R^2, RMSE, P_1, P_2 for our *elite* league dataset, split by season.

Season	R^2	RMSE	P_1	P_2
05/06	0.84576	0.10689	0.37340	0.58823
06/07	0.80332	0.12249	0.36796	0.59849
07/08	0.82067	0.10773	0.37081	0.59307
08/09	0.89556	0.08401	0.37495	0.58519
09/10	0.94893	0.06100	0.38075	0.57518
10/11	0.89620	0.08000	0.36912	0.59602
11/12	0.96151	0.04783	0.37626	0.58291
12/13	0.90565	0.08169	0.37919	0.57617
13/14	0.97634	0.03937	0.38644	0.56293
14/15	0.81375	0.12699	0.37990	0.57565
15/16	0.98860	0.02814	0.37731	0.58059
16/17	0.97868	0.04035	0.39273	0.55283
17/18	0.99342	0.02136	0.39202	0.55363
18/19	0.78709	0.13288	0.38508	0.56514
19/20	0.98955	0.02433	0.37715	0.57974

The figure clearly shows that accuracy is improving over time, for each variable: R^2 and P_1 are increasing; RMSE and P_2 are decreasing. This is likely due to advances in the betting models used by bookmakers, and increased information available.

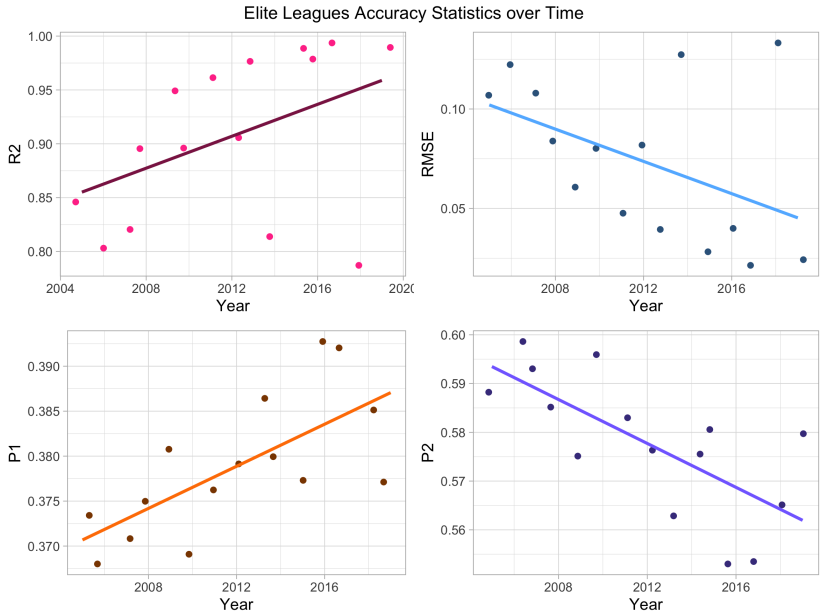

Figure 2.7: The variation of R^2, RMSE, P_1 and P_2 over time.

2.8 Principal Component Analysis

In this final analysis of our first chapter, we will conduct a pair of principal components analyses (PCA): firstly, by league; secondly, by season. PCA is a tool used in multivariate analysis to simplify the number of variables a data set has (Wold, Esbensen, and Geladi, 1987): we look to find "lines and planes of closest fit to systems of points in space" (Pearson, 1901). Graphically we can think of PCA as setting a 'new' pair of axes to a scatter plot: the 'new' x axis being on the line of best fit and the 'new' y axis being on the line of worst fit.

2.8.1 By-League Principal Component Analysis

To conduct our first PCA, we three new variables: *Imbalance*, *Level of Attack* and *Predictive Accuracy*. The imbalance variable, found for each country c using Equation 2.9, is a scaled average of the NAMSI N, inverse-κ (we take the inverse so high scores in all indicate higher imbalance), and Gini coefficient G (these variables were explained in Section 2.6) (Goossens, 2005).

$$\text{Imbalance}_c = \frac{1}{3}\left[\left(\frac{c_N - \mu_N}{\sigma_N}\right) + \left(\frac{c_\eta - \mu_\eta}{\sigma_\eta}\right) + \left(\frac{c_G - \mu_G}{\sigma_G}\right)\right] \tag{2.9}$$

The Level of Attack for each country c is found by Equation 2.10. For n seasons of data, we find the average number of shots per game (HS+AS) divided by the average number of goals per game (HG+AG). To compute this in R, we read in dataset from `football-data.co.uk` as we did in Section 2.2, using the code below.

$$\text{Attack}_c = \frac{1}{n}\sum_{s=1}^{n}\left(\frac{\mu_{(HS+AS),\ s}}{\mu_{(HG+AG),\ s}}\right) \tag{2.10}$$

(N.B., Our data for Portugal may al `football-data.co.uk` only has the shots data for the Premiera Liga from the 2017/18 season onwards.)

Similarly to imbalance, the Predictive Accuracy PA is a scaled average of our predictive variables: R^2, RMSE, P_1 and P_2 (we take the inverse of RMSE and P_2, say ζ and θ respectively), found by Equation 2.11.

$$\text{PA}_c = \frac{1}{4}\left[\left(\frac{c_{R^2}-\mu_{R^2}}{\sigma_{R^2}}\right)+\left(\frac{c_\zeta-\mu_\zeta}{\sigma_\zeta}\right)+\left(\frac{c_{P_1}-\mu_{P_1}}{\sigma_{P_1}}\right)+\left(\frac{c_\theta-\mu_\theta}{\sigma_\theta}\right)\right] \tag{2.11}$$

Once we have computed these values, we can run our PCA using the code below. This gives us our three principal components: PC1 can be considered a contrast between the imbalance and accuracy, against the style of play of a league; and PC2 a contrast of the predictive accuracy and the style of play, against the imbalance. The output from the code is in Table 2.11, where the *loadings* are the component values; the *importance* indicates 82% of the variation is explained by PC1; 17% by PC2 and only 1% by PC3. Finally, we create a screeplot, and scatter plot to visually inspect the PCA. These are in Figure 2.8.

```
pc.league <- matrix(c(imbalance, attack.league, predacc), ncol=3,
                    byrow=F)
colnames(pc.league) <- c("imbalance", "attack", "predacc")
rownames(pc.league) <- countries

league.model <- prcomp(pc.league)
league.model$rotation; summary(league.model)
```

Table 2.11: By-League PCA values.

Loadings	PC1	PC2	PC3
`imbalance`	0.7382378	0.4105132	-0.5352419
`attack`	-0.3206651	-0.4845176	-0.8138898
`predacc`	0.5934466	-0.7724776	0.2260519
Importance			
Standard deviation	1.166	0.5301	0.13284
Proportion of Variance	0.820	0.1694	0.01064
Cumulative Proportion	0.820	0.9894	1.00000

The PC1 v. PC2 figure interestingly groups our leagues. Portugal (with a high PC1 and PC2), Italy (neutral PC1, low PC2), Germany (low PC1, high PC2) are all in their own clusters, whilst Spain, England and France all have similar values for both components.

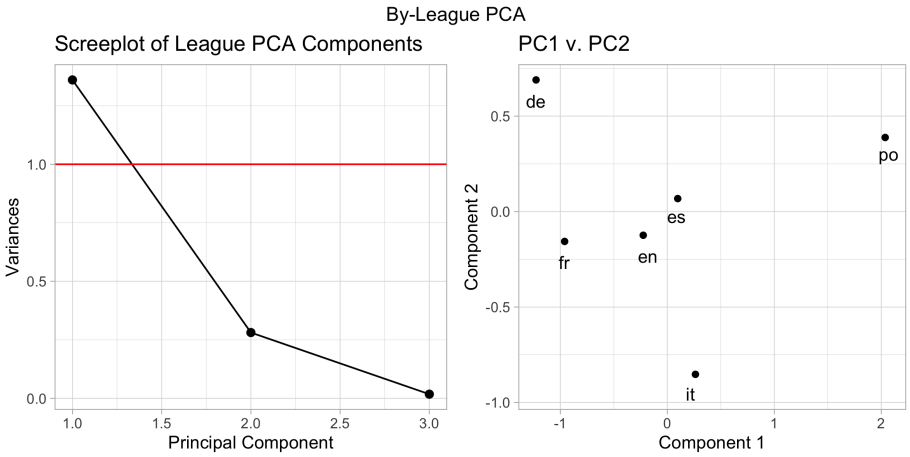

Figure 2.8: By-League PCA figures.

2.8.2 By-Season Principal Component Analysis

We conduct PCA with the following variables: R^2, inverse RMSE (ζ), P_1, inverse P_2 (θ), and the level of attack for that season, found by Equation 2.12 (where c represents a country; s a season.

$$\text{Attack}_s = \frac{1}{n}\sum_{c=1}^{n}\left(\frac{\mu_{(HS+AS),\ c}}{\mu_{(HG+AG),\ c}}\right) \tag{2.12}$$

Using R, we make use of the `prcomp` function to carry out our analysis. The full code, including the creation (and scaling) of a dataframe, is below; the outputs are in Table 2.12, as well as a screeplot and a plot of PC1 v. PC2 in Figure 2.9.

```
pc.season <- matrix(c(rsqu.season, (1/rmse.season), p1.season,
(1/p2.season), attack.season), ncol = 5, byrow=F)
colnames(pc.season) <- c("rsqu", "inv rmse", "p1", "inv p2", "attack")
rownames(pc.season) <- seasons
pc.season.sc <- scale(pc.season)

season.model <- prcomp(pc.season.sc)
summary(season.model); season.model$rotation
```

Which principal components do we keep?

From our importance of components output (Table 2.12) and the screeplot (Figure 2.9), we can see that the first two components account for 85.6% of the cumulative variance. On the screeplot, we add the line $y = 1$ to represent the Kaiser criterion (red), which, whilst arbitrary and not recommended as a *hard-and-fast* rule (Fabrigar et al., 1999), can be a good guide. This choice also follows the criterion of retaining eigenvalues (it can be shown the eigenvalues are equal to the variances (Alto, 2019)) greater than 0.7 times the average

Table 2.12: By-Season PCA values.

Importance of components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.8112	0.9998	0.7909	0.30632	0.02448
Proportion of Variance	0.6561	0.1999	0.1251	0.01877	0.00012
Cumulative Proportion	0.6561	0.8560	0.9811	0.99988	1.00000

Component rotations:

	PC1	PC2	PC3	PC4	PC5
R^2	0.409	-0.624	0.199	0.636	0.013
Inverse RMSE, ζ	0.453	-0.495	-0.241	-0.701	-0.010
P_1	0.489	0.364	0.360	-0.055	-0.704
Inverse P_2, θ	0.491	0.372	0.333	-0.070	0.710
Level of Attack	-0.383	-0.308	0.813	-0.310	0.014

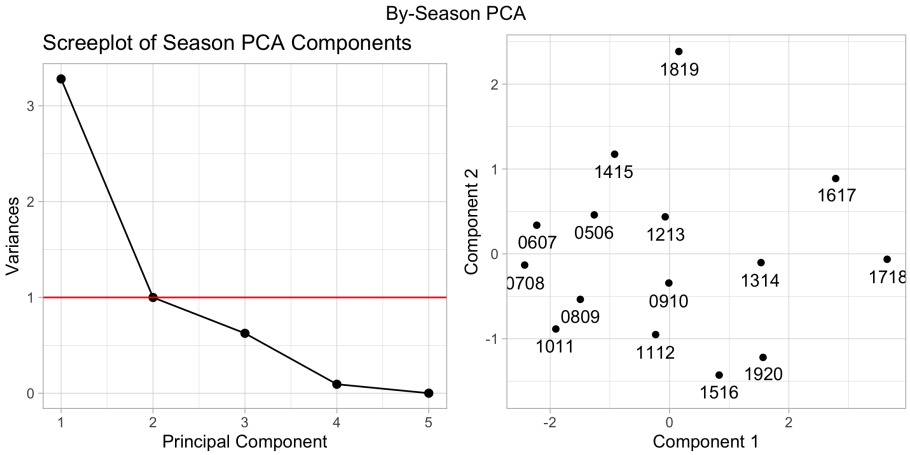

Figure 2.9: By-Season PCA figures.

eigenvalue (Jolliffe, 1972): in our case, this is 0.14, also recommending to keep PC1 and PC2.[13]

Interpreting our components

PC1 can be interpreted as a contrast between our predictive variables (R^2, ζ, P_1, θ), and the level of attack, with high values being awarded to seasons with high levels of accuracy, and/or lower levels of attacking football.

PC2 can be interpreted as a contrast between the P_1 and inverse P_2 values, and—mainly—R^2. Whilst both are statistics of accuracy, the two are not mutually agreeable: in Table 2.10, the 2018/19 season has the extreme-poor-performing R^2 and RMSE, but performs on-trend for P_1 and P_2 (Figure 2.7).

In the PC1 v. PC2 plot in Figure 2.9, we see the earlier years (2005/06 until

[13] `0.7*mean(c(0.6561,0.1999,0.1251,0.01877,0.00012))=0.1399986`

2012/13) are contained within the bottom-left (low PC1, low PC2) of the graph (bounded by $x = 0,\ y = 0.5$); in the bottom right, we have the 15/16, 19/20, 13/14 and 17/18 seasons, with low PC2, and high PC1: this indicates lower attacking football (high PC1), and high levels of accuracy, especially the R^2 and RMSE measures, indicating both a trend of accuracy increasing over time (as mentioned in Section 2.7), and of more recent football seasons being more attacking. This is an area for future research, perhaps using a larger dataset and more advanced measures of attacking styles of play.

2.9 Conclusion

In this chapter, we have shown that the levels of bookmaker accuracy are high in the home win and away win market: in each of our *elite* European leagues, the coefficient of determination R^2 was above 95%, with the RMSE below 0.055. From our scatter plot (Figure 2.6), we saw the linear models from home and away wins have a much lower standard error (and therefore a narrower confidence interval) than the model for draws. This, along with the R^2 and RMSE for each result (Table 2.5), indicates whilst bookmakers enjoy high accuracy with clear results, their predictions for draws are poor, and have large room for improvement.

In addition, we have shown that the accuracy is impacted by the competitive balance in each league: bookmakers perform better in countries in unbalanced leagues, such as Portugal's Primiera Liga and Italy's Serie A, than in balanced leagues, such as Germany's Bundesliga and France's Ligue Une.

We have also located room for future research: there may be a relationship between home advantage and bookmaker's accuracy of away games (perhaps this differs between leagues, too?), and there may also be a relationship between the distribution of bookmaker's odds, and the style of play, in certain leagues, such as the defensive Serie A displaying a unimodal distribution and the offensive La Liga displaying a trimodal distribution. This difference in the style of play may also impact bookmaker accuracy.

In Chapter 3, we investigate the accuracy of bookmaker's odds in the English and Scottish leagues, across multiple levels (rather than just the *elite* leagues) and betting markets (rather than just the 1X2 market).

Chapter 3

Assessing the accuracy of betting odds in the English and Scottish football league pyramids, from 2005 to 2020.

In this chapter, we aim to assess the accuracy of betting odds across multiple and tiers in the football pyramid. To do this, we will use the English and Scottish football leagues, and the Asian Handicap (AH) and Under/Over 2.5 Goals (UO) markets, in addition to the 1X2 market as in Chapter 2. We choose the English and Scottish leagues due to `football-data.co.uk` having data available for five and four divisions in the respective systems: the maximum in any other country is two (Buchdahl, n.d.[c]). We will utilise many of the same techniques, including conducting exploratory data analysis (EDA), correlation analysis, and assessing the predictive power. We will also assess the OVERROUND across the markets and leagues, a measure of the bookmaker commission.

3.1 Exploratory Data Analysis

Reading and cleaning our data

As we have already checked our data source (Section 2.1), we will proceed directly with our EDA. We again use a `for` loop to read the data directly from `football-data.co.uk`; we use a similar code as in Section 2.2.

Once the data is imported, we define the `level` of the league: we define Level 1 as the English Premier League (EPL) and Championship, and the Scottish Premier League (SPL). The EPL and the SPL are the top-tier leagues in their pyramids; the Championship's average revenue per club is €33 per club, per year (€13 on average per year more than the SPL) (Ajadi et al., 2020). We define Level 2 as the remaining fully professional leagues in the pyramids: these

are the English Leagues One and Two, and the Scottish Division One. Finally, we define Level 3 as the leagues with semi-professional sides in: the English Conference and Scottish Divisions Two and Three. The code to import the data is below: in addition to previous variables, we read in the Asian Handicap (AH) and Under/Over 2.5 Goals market variables (see Table 1.3).

After reading in our data, and adding the `level` using a `for` loop, we find the pre- and post-normalised probabilities, using the same method as in Section 1.3, Equation 1.1. There are a couple of major mistakes in `football-data.co.uk`'s files with handicaps recorded, which we fix below.

```
#A few important notes:
#Rangers had a -275 goal handicap v. East Fife; assume this meant -2.75:
ensco$HomeHandicap[ensco$HomeTeam=="Rangers" &
                     ensco$Date=="11/01/14"] <- -2.75
#Hamilton had a 12.5 goal handicap v. Rangers; assume this meant 1.25:
ensco$HomeHandicap[ensco$HomeTeam=="Hamilton" &
                     ensco$Date=="25/10/08"] <- 1.25
```

Winning Probabilities

We find the correct probability (that is, the bookmaker consensus probability of the event that occurred) using a set of `for` loops: this is simple for the 1X2 market (using the `FTR`) column) and the UO market (we create a new column `TotGoals = FTHG + FTAG` and assess whether it is under or over 2.5), but requires more thought for the AH market.

First, we find the AH goal difference (or 'gap') between the two sides. Full-goal handicaps can result in a home or away win, or a draw (at which point the bet is voided (bet365, n.d.)). A half-goal handicap can only result in a home or away win. A quarter-goal handicap (E.G. 0.75), however, can result in a *half-win* for the bettor: half of the stake is assigned to the nearest (in a number-line sense) half-handicap (0.5), and half to the nearest integer (1). If only one bet wins, the bettor wins on half their stake. We use the R code below to encode this.

```
ensco$ah.gap <- with(ensco, FTHG.ah - FTAG); ensco$ah.res <- NULL
for (n in 1:N){
    if (ensco$ah.gap[n]<(-0.25)){ensco$ah.res[n]<-"aw"}
    else if (ensco$ah.gap[n]==(-0.25)){ensco$ah.res[n]<-"hfaw"}
    else if (ensco$ah.gap[n]==0){ensco$ah.res[n]<-"vo"}
    else if (ensco$ah.gap[n]==0.25){ensco$ah.res[n]<-"hfhm"}
    else if (ensco$ah.gap[n]>0.25){ensco$ah.res[n]<-"hm"}
    else{}
}
```

Basic Calculations

As with Section 2.2, our first step is to compute the bookmaker consensus mean probabilities (and corresponding standard deviations), and compare these to

the observed probabilities: we choose to sort these by level. These are shown in Table 3.1.

Table 3.1: EDA calculations

1X2 Market	*Level 1*	*Level 2*	*Level 3*
Mean $\mathbb{P}_{\text{cons}}$(Home Win)	0.4366	0.4257	0.4292
$\mathbb{P}_{\text{obs}}$(Home Win)	0.4438	0.4256	0.4352
Standard Deviation (Home Win 1x2)	0.1507	0.1044	0.1238
Mean $\mathbb{P}_{\text{cons}}$(Draw)	0.2644	0.2725	0.2613
$\mathbb{P}_{\text{obs}}$(Draw)	0.2607	0.2695	0.2406
Standard Deviation (Draw 1x2)	0.0369	0.0195	0.0243
Mean $\mathbb{P}_{\text{cons}}$(Away Win)	0.2990	0.3018	0.3095
$\mathbb{P}_{\text{obs}}$(Away Win)	0.2955	0.3049	0.3242
Standard Deviation (Away Win 1x2)	0.1366	0.0952	0.1139

Under/Over 2.5 Goals Market	*Level 1*	*Level 2*	*Level 3*
Mean $\mathbb{P}_{\text{cons}}$(Under)	0.5080	0.5121	0.4756
$\mathbb{P}_{\text{obs}}$(Under)	0.5051	0.5184	0.4755
Mean $\mathbb{P}_{\text{cons}}$(Over)	0.4920	0.4879	0.5244
$\mathbb{P}_{\text{obs}}$(Over)	0.4949	0.4816	0.5245
Standard Deviation (Under/Over Market)	0.0568	0.0380	0.0504

Asian Handicap Market	*Level 1*	*Level 2*	*Level 3*
Mean $\mathbb{P}_{\text{cons}}$(AH Home Win)	0.5109	0.5046	0.5015
$\mathbb{P}_{\text{obs}}$(AH Home Win)	0.4011	0.4005	0.4000
$\mathbb{P}_{\text{obs}}$(AH Half-Home Win)	0.0434	0.0513	0.0583
Mean $\mathbb{P}_{\text{cons}}$(AH Away Win)	0.4891	0.4954	0.4985
$\mathbb{P}_{\text{obs}}$(AH Away Win)	0.3779	0.3920	0.4087
$\mathbb{P}_{\text{obs}}$(AH Half-Away Win)	0.0620	0.0741	0.0659
$\mathbb{P}_{\text{obs}}$(Void Bets)	0.1157	0.0821	0.0671
Standard Deviation (Asian Handicap Market)	0.0678	0.0524	0.0462

To take this table apart, we first look into the 1X2 market: as with in Chapter 2, we notice the standard deviation, for all levels, is far lower for Draws than for either Home or Away Wins, indicating a consistent lack of variation in the odds offered for draws. It is noticeable, too, that the standard deviations for all three outcomes is lowest for Level 2 (rather than Level 3, as one would expect, due to the amount of, or lack thereof, information available). For all three outcomes, across all three levels, the consensus probability is very close to the observed probability; the clearest examples of this are the Level 2 Home Win probabilities: 0.4257 (consensus) and 0.4256 (observed), and the Level 1 Away Win probabilities: 0.2990 (consensus) and 0.2955 (observed).

As, for the UO and AH markets, bookmakers offer only two options, the standard deviations are equal. For the UO market, we notice the standard deviations are very low (especially in Level 2, where $\sigma = 0.0380$), and the consensus probabilities are *highly* accurate: in Level 3, the two differ by 0.0001.

The mean number of goals across all matches in our data set is 2.66, so it is no surprise the means are around 0.5. It is likely the bookmakers know the mean number of goals in each level (or league), and set their odds accordingly.

Lastly, we consider the AH market. The numbers aren't as easy to consider due to the half-wins and voided bets: of the full Home and Away Wins, 50.57% were Home Wins: by level, the proportions of full Home Wins are 0.5149, 0.5054 and 0.4946 for Levels 1, 2, and 3 respectively. Again, these are all remarkably close to the bookmaker consensus means. It is, as with the UO market, unsurprising these values are near 0.5: the AH market is designed to give a handicap in favour of the poorer side in the form of a goal deficit to the superior side (Constantinou, 2020).

Visual Analysis

To help us visualise these figures, we will produce a number of plots, beginning with density plots; these are in Figure 3.1.

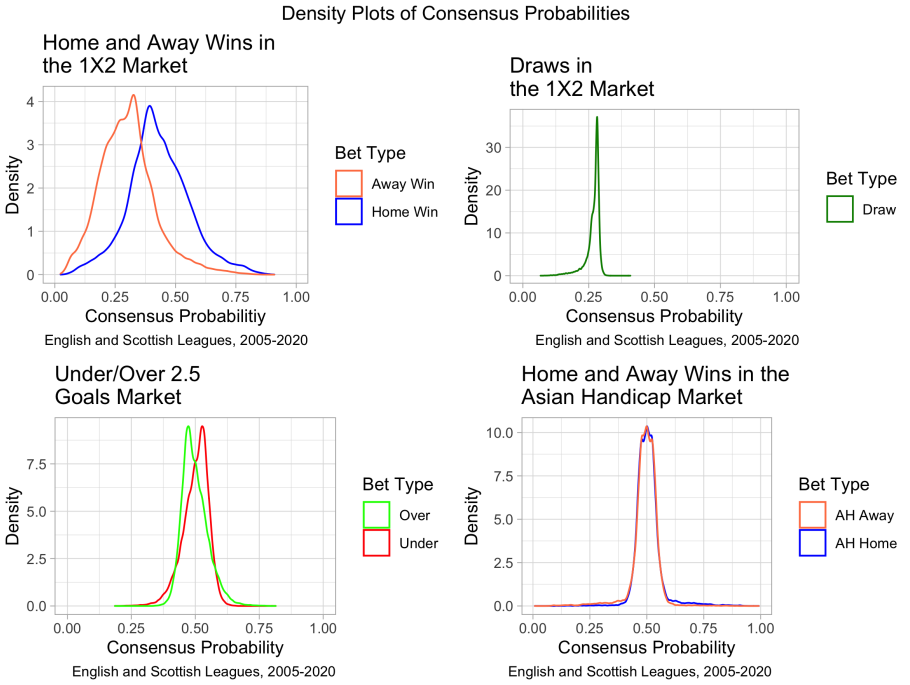

Figure 3.1: Density plots of the consensus probabilities offered in the 1X2, UO and AH markets.

The 1X2 densities are similar to those we saw in Figure 2.3, with the Home Win curve being symmetrically-distributed about the mean at around 0.45; the Away Win curve having a positive skew; and the Draw curve having a negative skew, with a sharp drop (and few games) after the MODE at ≈ 0.27, with low overall variance. As expected, due to its nature of encouraging parity, the AH

density curves have low variation, and are have their mean at 0.5. The Over 2.5 Goals curve has a positive skew, with the mode below the mean. It has low variance, and is—as expected with only two outcomes—a reflection (Weisstein, n.d.) of the Under 2.5 Goals density about the line $x = 0.5$.

A further visual aid we can produce is an investigation into the handicaps offered by bookmakers. We would expect teams with a higher $\mathbb{P}_{\text{cons}}$(Home Win) would have a more negative handicap, and vice-versa. To visualise this, we create two plots: the Handicap offered versus the consensus Home Win probability, which we will split by the level, before adding the consensus Away Win probability, ensuring the Handicap holds for both. These are shown in Figures 3.2 and 3.3.

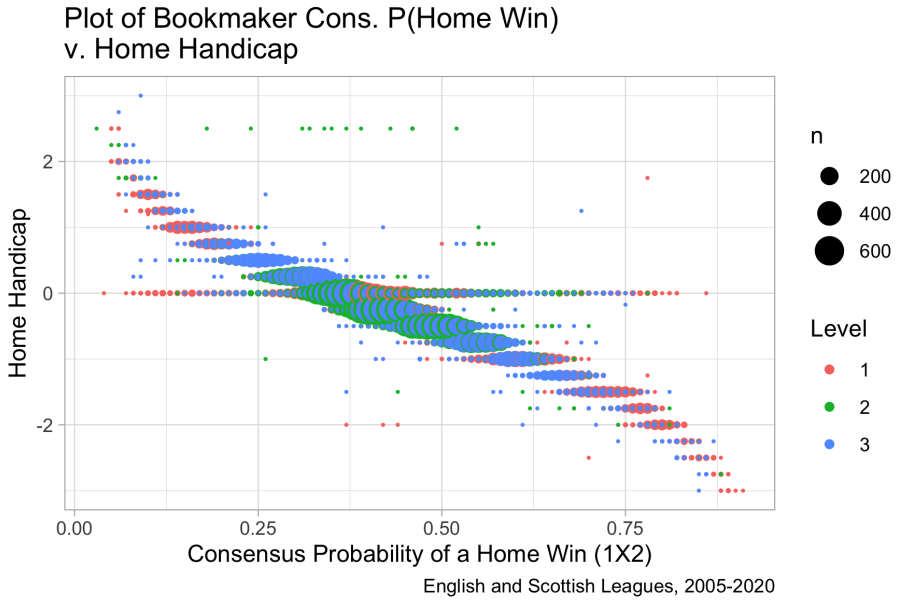

Figure 3.2: The AH Handicap versus the consensus probability of a Home Win, split by level.

Both figures indicate what we would expect, with a trend suggesting high Win probabilities equate to a more negative handicap. The figures infer a non-linear trend, with small increases in probability equating to a larger change in handicap at the extremes, though our sample sizes here are small. There are a number of matches, at all levels, with 0 handicap across the range of probabilities, indicating bookmakers are often happy without assigning a handicap, even if a match is highly in favour of one side.

As in Figure 2.5 in Section 2.2, we can produce a tile plot, to see if the correct market probabilities are greater for 'more convincing' matches, those with a greater final goal difference. We look into this for both the 1X2 and Under/Over 2.5 Goals market. We would expect similar findings for the former as in Figure 2.5, and we expect the latter to show the same trend, if the bookmakers are accurate (as we know the variance of consensus probabilities is low

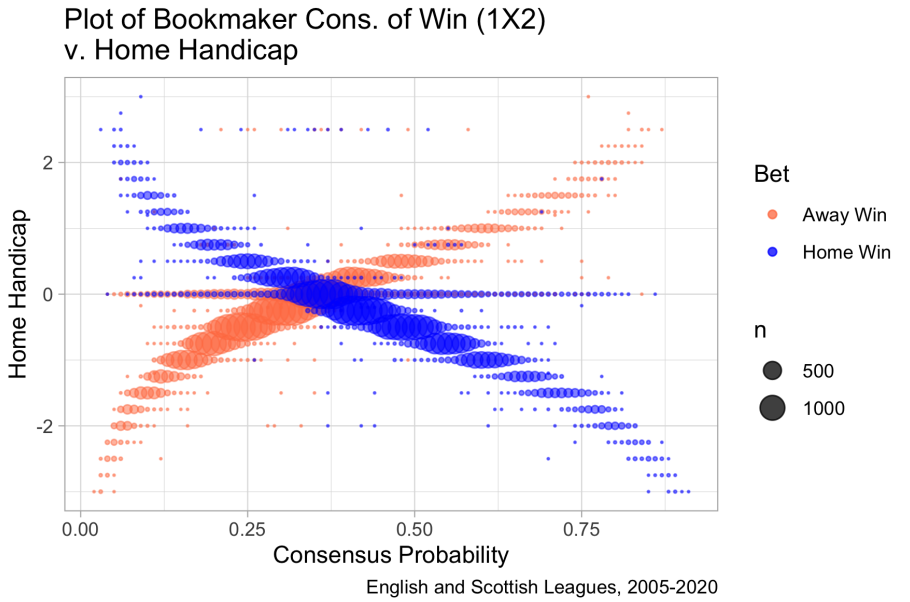

Figure 3.3: The AH Handicap versus the consensus probabilities of both a Home Win and Away Win.

(Table 3.1), we use a different `ggplot2` palette). These plots are in Figures 3.4 and 3.5; the bin sizes[1] are shown in Table 3.2.

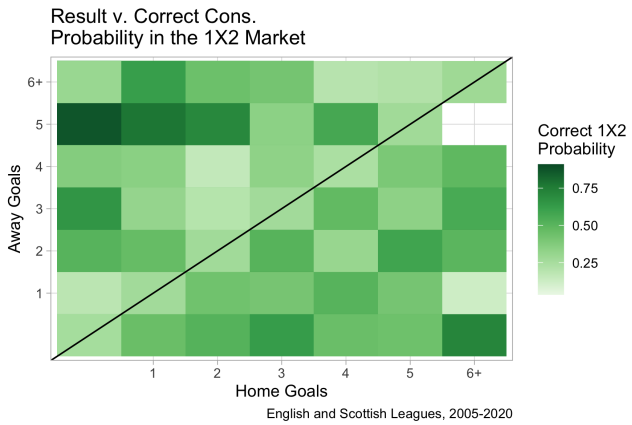

Figure 3.4: Tile plot of the correct 1X2 probability versus the full time result.

From Figure 3.4, it is observable that more convincing results have a higher correct probability. The highest scoring tiles are highly-convincing away wins (0-5, 1-5, 2-5, 1-6+ and 0-3 are five examples of the most noticeable such tiles)

[1]In this dataset, our highest scoring draw finished 6-6, between Motherwell and Hibernian, 05/05/10, in the Scottish Premiership, so we group 6+ as our highest goal category.

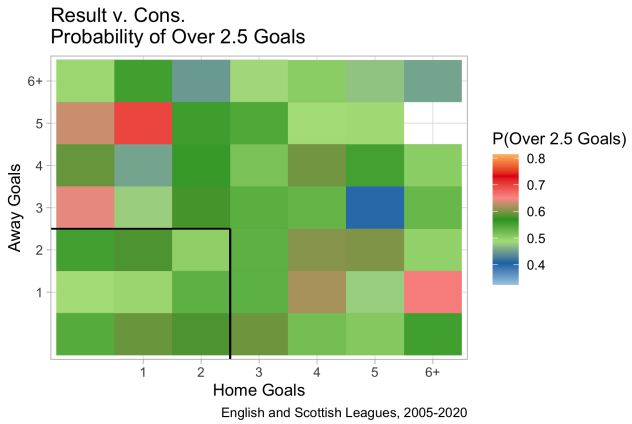

Figure 3.5: Tile plot of the correct UO probability versus the full time result.

Table 3.2: The Bin Size for each tile of Figures 3.4 and 3.5

Away Goals								
	6+	48	38	16	12	1	1	1
	5	115	133	74	23	18	6	0
	4	407	446	268	144	54	19	5
	3	1085	1337	1003	508	167	47	7
	2	2388	3530	2628	1242	411	125	44
	1	4091	6012	4372	1993	763	245	119
	0	3590	5009	3761	1954	782	304	132
		0	1	2	3	4	5	6+
		Home Goals						

and highly-convincing home wins (6+-0, 3-0 and 5-2), though the former is more clear. It is hard to draw meaningful conclusions from the highest scoring matches (the top right of the plot) due to low sample sizes, but at the bottom left (lowest scoring matches, where n_{bin} is high), the pattern holds well.

In Figure 3.5, we add a black line at 2.5 goals: the nine tiles within these lines are where Under is the correct bet: there are not many permutation, but the observed probability (for the entire dataset) of this, P_{obs}(Under 2.5 Goals) = 0.5023, so the tiles outside this will be more telling. As there is low variation in the odds offered, we use a different `ggplot2` palette: the red squares have a higher consensus probability of Over 2.5 Goals; the blue squares are lower.

The 'most-red' tiles (1-5, 6-1 and 0-3, $n_{1-5} = 133$, $n_{6-1} = 119$ and $n_{0-3} = 1085$) generally have more goals: the 0-3 result is the exception. We can assume these results are where the away team is the favourite, and it appears bookmakers predict the side to score more: an area to investigate is whether strong favourites play more conservatively when away from home than bookmakers expect. The trend does not follow strictly, either. The 5-3 Home Win tile ($n_{5-3} = 47$)has the lowest correct probability, at around 0.4, despite eight goals being scored in these matches. Other high-scoring tiles with low correct probabilities include the 1-4, 2-6+, and 6+-6+ draws: each of these has a low

n, though. The most striking conclusion this tile plot gives us is the *lack* of a pattern: there is no obvious feature, suggesting—as with the draws in the 1X2 market—bookmakers struggle at placing reliable odds in the UO market.

Our final two plots in this analysis are *count plots*: a way of visualising two discrete variables (Plotly.com, n.d.). The first, Figure 3.6, is a plot of the mean home handicap offered by bookmakers against the consensus probability of over 2.5 goals in a match. We would expect the largest magnitude handicaps are for more convincing matches, and thus will have a greater probability of over 2.5 goals. The second, Figure 3.7, is a plot of the expected goal difference (the handicap) against the observed goal difference. If the handicaps offered were perfectly accurate, this would be a linear line on $y = x$.

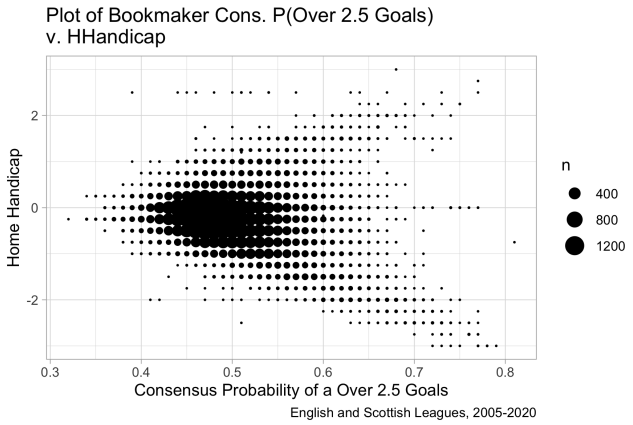

Figure 3.6: Count plot of the home handicap offered versus the consensus probability of over 2.5 goals.

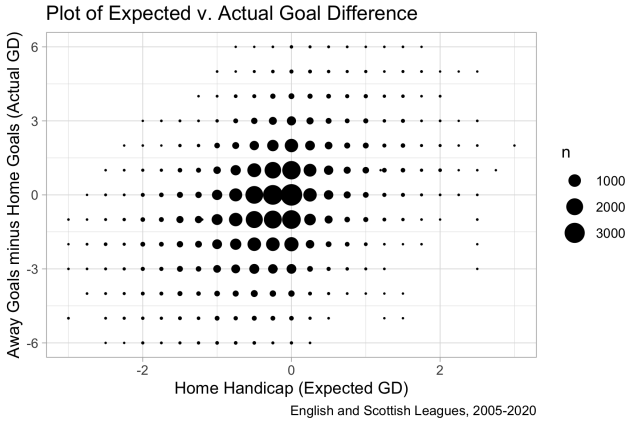

Figure 3.7: The actual v. observed goal difference.

In both plots, the expected general trend is observed, though neither plot

is as convincing as say Figure 3.2; they have a lot of NOISE. As a Home Win is more likely (Table 3.1), it is unsurprising that there is a greater amount of handicaps against the home side.

From all of these figures, we can see that there is generally less accuracy from the bookmakers in the Under/Over market (see Figure 3.5) compared to the Asian Handicap and 1X2 markets. We have evidence to suggest the handicap offered is accurate, but this will require further investigation.

3.2 Correlation Analysis

We conduct a correlation analysis in a similar fashion to Chapter 2. For all three markets, we first *bin* (we choose $n = 50$ across all markets) our data, and set the levels of each bin using the R `cut` and `tapply` commands. We find the observed probabilities of each bin using the `prop.table` function: for the Asian Handicap market, we will only consider the full wins for ease. As the Asian Handicap and Under/Over 2.5 Goals markets are binary, we only need to create one linear model. The coefficients of our six final models to predict the observed probability $\mathfrak{O}$ from the consensus probability $\mathfrak{C}$ are shown in Table 3.3.

Table 3.3: Linear models to predict the observed outcome $\mathfrak{O}$ from a given bookmaker consensus probability $\mathfrak{C}$ using the English/Scottish Data.

Market	*Intercept*	*Slope*
1X2 Home Win	-0.0153	1.0637
1X2 Draw	-0.1797	1.9814
1X2 Away Win	-0.0111	1.0462
1X2 Overall	-0.0089	1.0642
Under/Over	0.0358	0.8463
Asian Handicap	-0.0419	0.8691

For example, if we were to know the bookmaker consensus probabilities for an Asian Handicap bet and a Home Win 1X2 bet were both 0.6, we can predict the respective observed probabilities being 0.48 and 0.62. The models aren't much help for this use: they can be used instead to find values for the coefficient of determination R^2 and root mean squared error RMSE. These are in Table 3.4.

Table 3.4: Values for R^2 and RMSE for the markets of interested, based on the models in Table 3.3

1X2 Market	*Home Win*	*Draw*	*Away Win*
R^2	0.9917491	0.5176747	0.97147833
RMSE	0.0246581	0.1601749	0.04393204
	Under/Over	**Asian Handicap**	
R^2	0.3868131	0.89587870	
RMSE	0.1478373	0.07556164	

The values clearly indicate the poor levels of accuracy in the Under/Over 2.5 Goals market, and for the bookmakers odds for a draw, with high levels of accuracy in the Asian Handicap market (low RMSE, and $\approx 90\%$ R^2 value). Our 1X2 market findings are consistent with our findings outlined in Chapter 2.

To visualise these findings, we produce plots. These are shown in Figures 3.8, 3.9, and 3.10. The figures also contain the linear model and corresponding 95% confidence interval, as defined in Appendix A.1, Definition 3.

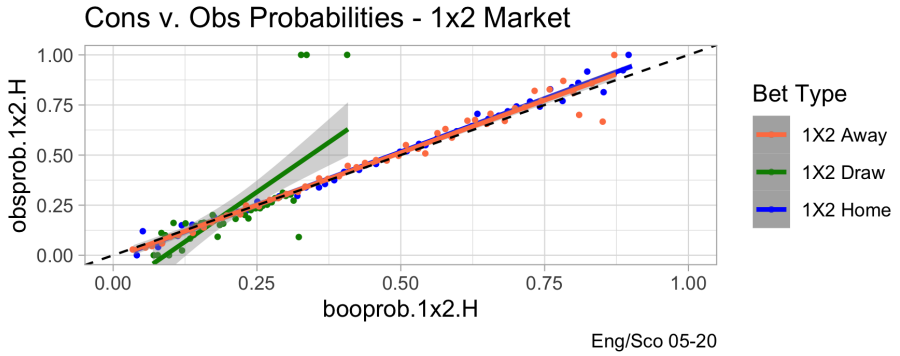

Figure 3.8: Consensus v. Observed probabilities on the English & Scottish data, 1X2 market.

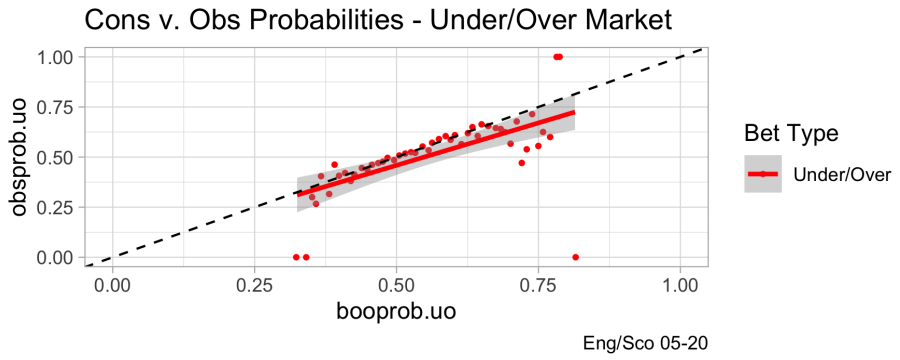

Figure 3.9: Consensus v. Observed probabilities on the English & Scottish data, Under/Over 2.5 Goals market.

The figures support our conclusions from the R^2 and RMSE values. Figure 3.8 has a near-perfect correlation for the Home Wins and Away Wins, corresponding to the respective R^2 values: 0.992 and 0.971. The Draws are inaccurate, with a large confidence interval, implying a large standard error (in fact, the SE is 0.3063). A quick look at Figure 3.9 suggests relatively high levels of accuracy; our R^2 and RMSE values may be due to high values of leverage at the extremities, emphasised by the low levels of variance (3.1) in the odds. Figure 3.10 shows high levels of accuracy in the Asian Handicap market, though

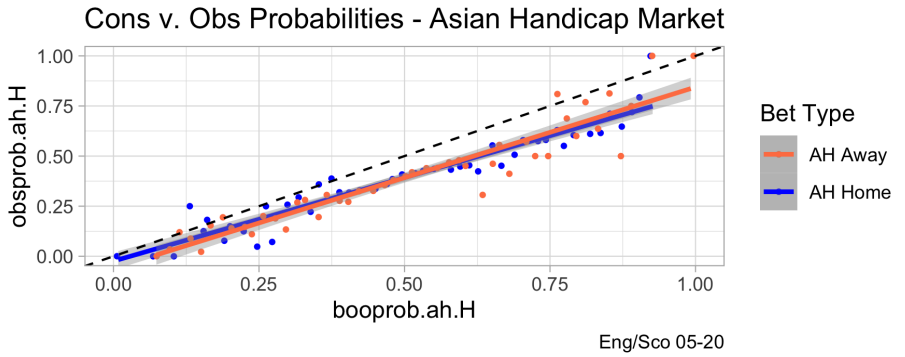

Figure 3.10: Consensus v. Observed probabilities on the English & Scottish data, Asian Handicap market.

both lines are significantly[2] below the $y = x$ (dotted) line. To further analyse these models, we consider each level separately.

3.3 Comparing Levels

Using a `for` loop in R, we can complete the same process as above, to find the R^2 and RMSE for each level, in each market. We will also, for the 1X2 market, find the values of P_1 and P_2. As in Section 2.5, we weight the bins, with more value being added to the Home Win and Away Win bets ($n = 35$) than to the Draw bets ($n = 15$) in the 1X2 market. For the Under/Over 2.5 Goals and Asian Handicap markets, our bin size $n = 35$. These figures are shown in Figures 3.11, 3.12, and 3.13, with values of statistical accuracy in Table 3.5.

Table 3.5: R^2 and RMSE values for the 1X2, UO and AH markets across all three levels, and P_1 and P_2 for the 1X2 market.

Level	1	2	3
Number of matches n	17317	18913	13248
1X2 Market			
R^2	0.99035	0.75579	0.91903
RMSE	0.02611	0.12914	0.07094
P_1	0.36767	0.35137	0.35902
P_2	0.59823	0.62984	0.61429
Under/Over 2.5 Goals Market			
R^2	0.63719	0.53157	0.02919
RMSE	0.10474	0.12414	0.12980
Asian Handicap Goals Market			
R^2	0.81714	0.76950	0.46815
RMSE	0.09766	0.10212	0.19181

[2]No part of the 95% confidence interval lies above the dotted line.

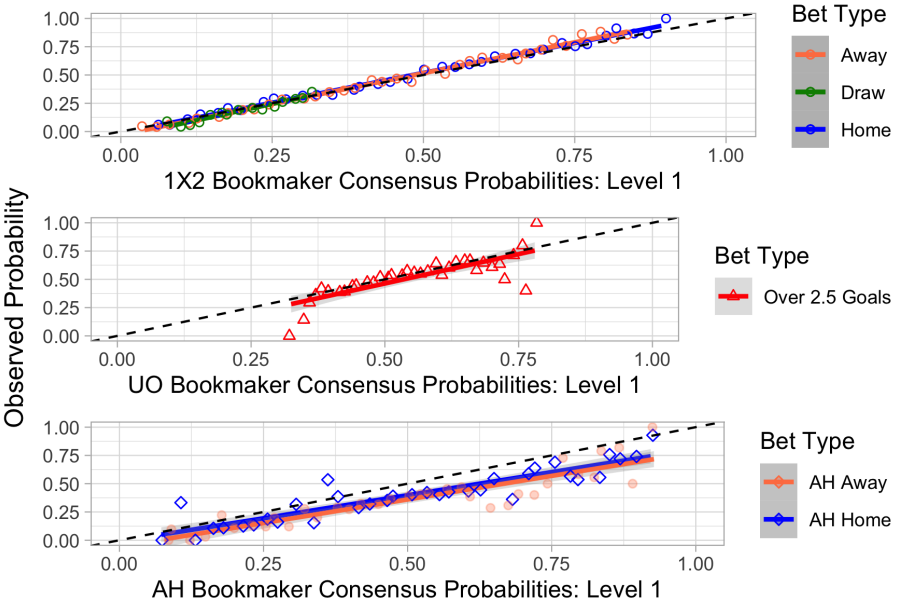

Figure 3.11: Plots of Consensus v. Observed Probabilities in the English & Scottish Dataset, Level 1.

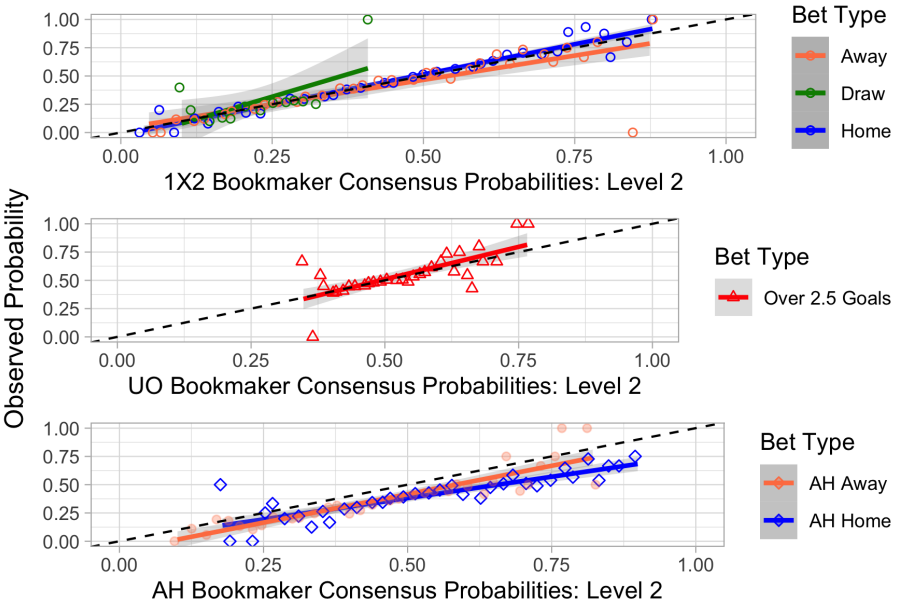

Figure 3.12: Plots of Consensus v. Observed Probabilities in the English & Scottish Dataset, Level 2.

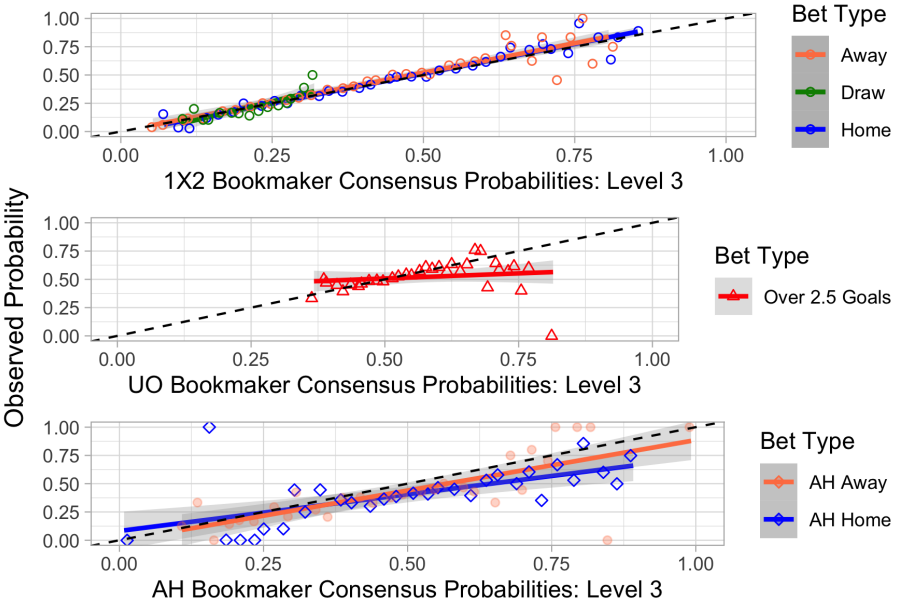

Figure 3.13: Plots of Consensus v. Observed Probabilities in the English & Scottish Dataset, Level 3.

The figures and values show an unexpected result. We saw Level 2 had lower variation than Level 3 in Section 3.1: the bookmaker's performance is also worse in Level 2 than in Level 3 (we would expect Level 2, which has more information available, to perform better). The confidence intervals are larger in the 1X2 market, and whilst the LINE OF BEST FIT seems better for the Under/Over 2.5 Goals market ($R^2_{L2} = 0.53$, $R^2_{L3} = 0.03$), we can observe more OUTLIERS in Level 2 than in 3 (RMSE of Level 2 = 0.124, RMSE of Level 3 = 0.130: much closer than the R^2 values suggest). A further investigation into the outliers using the principals of statistical leverage would be required to reach a formal conclusion. Level 2 does, however, appear to perform better in the Asian Handicap market. For all three markets, as we would predict, Level 1 performs significantly better than the other two. Across all three levels, bookmakers perform worst in the Under/Over 2.5 Goals market consistently, with the lowest R^2 for each level being for this market.

For the 1X2 market, we have found values of P_1 and P_2 as defined in Section 2.4. Better predictive performance is indicated with higher values of P_1 and lower values of P_2 (Owen, 2009). The P values for Level 1 are $P_1 = 0.36767$, which is lower all *elite* leagues with the exception of the French Ligue Une, and $P_2 = 0.59823$, higher than all *elite* leagues except Ligue Une. All four of the P values for Levels 2 and 3 indicate worse performance than all *elite* leagues. Comparing Levels 2 and 3, we find $P_{1,\,L3} > P_{1,\,L2}$ and $P_{2,\,L3} < P_{2,\,L2}$, indicating that the predictive performance in Level 2 is better (for the 1X2 market) in Level 3 (the lower leagues containing semi-professional teams) than Level 2 (lower leagues containing only professional sides). This is an unexpected result, but could be due to several reasons, such as competitive balance, the effect of the

FAVOURITE-LONGSHOT BIAS, a smaller sample ($n_{L3} = 13248 < 18913 = n_{L2}$), or simply: bookmakers finding the matches at this level harder to model. In Level 3, some sides will be fully-professional (for example, those who have recently been relegated, or who are attempting to reach promotion to higher standards) whereas others will be semi-professional (those who have recently been promoted from *regional* leagues). The gulf of quality between the two is likely to be larger than between the best and worst teams in Level 2. Further research into this may find reasoning behind this result.

3.4 The Overround

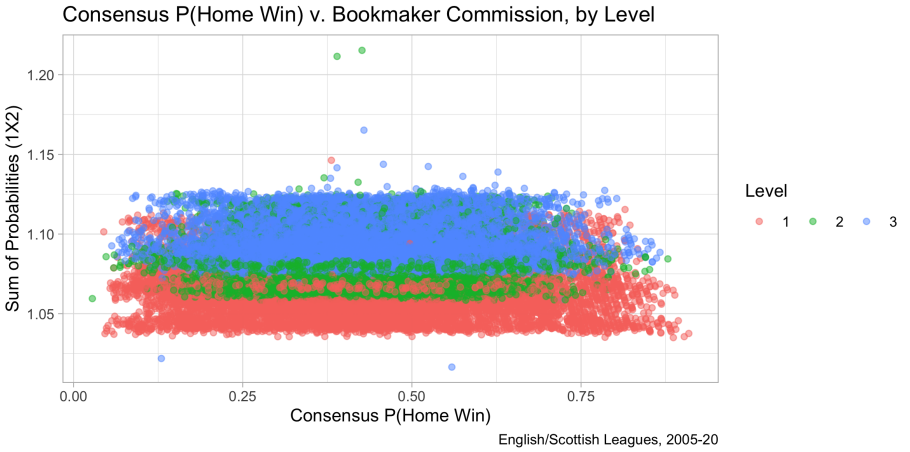

Figure 3.14: The sum of the $\mathbb{P}_{\text{cons}}$(1X2 Outcomes) v. $\mathbb{P}_{\text{cons}}$(1X2 Home Win), split by Level.

The *overround* η for a single match is defined in Equation 3.1. It is a measure of bookmaker commission: strictly, it is the profit made by the bookmaker when a bettor places a bet of $\beta_i = \frac{1}{1+O_i}$ on all n outcomes (guaranteeing a unit return) (Henery, 1999). An overround of 0.05 (the sum of probabilities = 1.05) represents a 5% bookmaker commission for that match, in that market. The overround is not constant, either, and may change among "matches, bookmakers and over time" (Angelini and De Angelis, 2019). (This is why we have normalised our odds prior to this). To assess where the bookmakers earn their commission, we will consider basic calculations and jitter/scatter plots of the overround in a given market, grouped by certain variables (level and season). The values are in Table 3.6; the plots are shown in Figures 3.14 to 3.19.

$$\eta = \sum_{i=1}^{n} \left[\frac{1}{O_i} \right] - 1 = \sum_{i=1}^{n} [p_i] - 1 \tag{3.1}$$

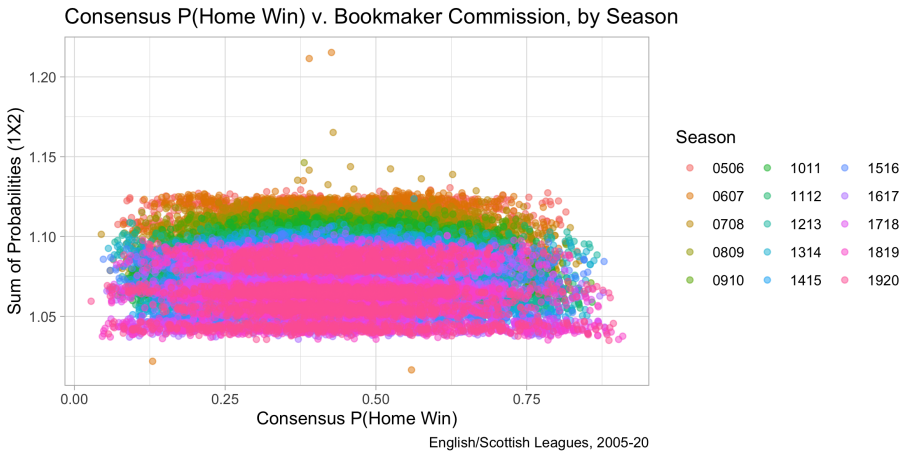

Figure 3.15: The sum of the $\mathbb{P}_{\text{cons}}$(1X2 Outcomes) v. $\mathbb{P}_{\text{cons}}$(1X2 Home Win), split by Season.

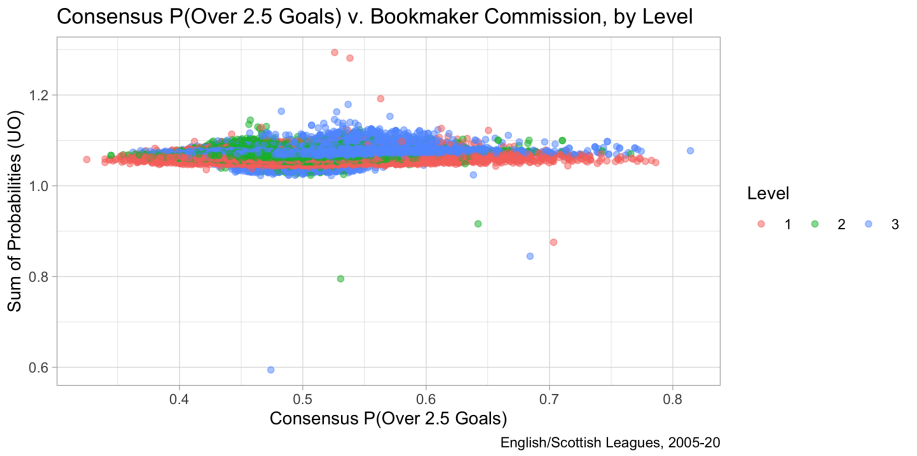

Figure 3.16: The sum of the $\mathbb{P}_{\text{cons}}$(UO Outcomes) v. $\mathbb{P}_{\text{cons}}$(Over 2.5 Goals), split by Level.

Overround by Level

Figures 3.14, 3.16, and 3.18 show the overround of each match, grouped by level, with red, green, and blue points for Levels 1, 2, and 3 respectively. In the 1X2 market, we can notice Level 1 has more varied overround: Level 3 varies from approximately 7% to 12%; Level 1 from 3% to 11%.[3]

In the UO market, the overround seems to be more evenly distributed, with more variation (across all levels) for matches when the odds nearer 2

[3]It is worth noting the matches are plotted first in order of time, then in order of league, so more recent matches in Level 3 are the datapoints added last. We can partially resolve this by altering the `alpha` (opacity) of the points. We choose this to be 50%.

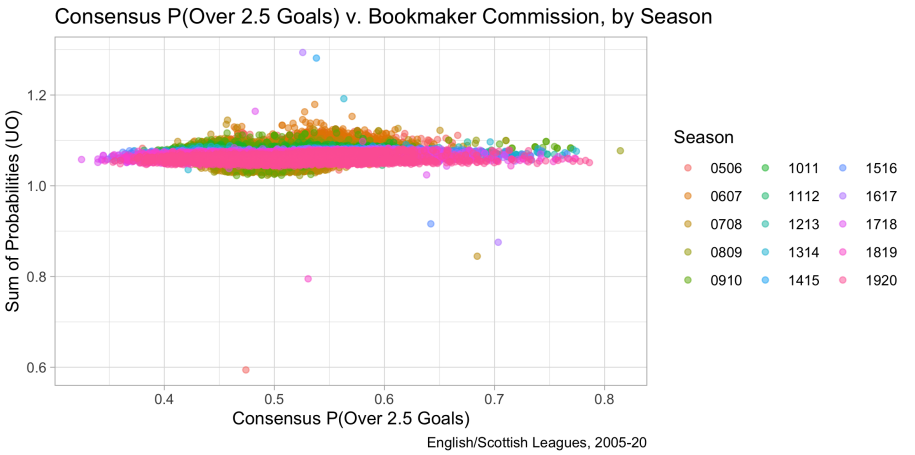

Figure 3.17: The sum of the $\mathbb{P}_{\text{cons}}$(UO Outcomes) v. $\mathbb{P}_{\text{cons}}$(Over 2.5 Goals), split by Season.

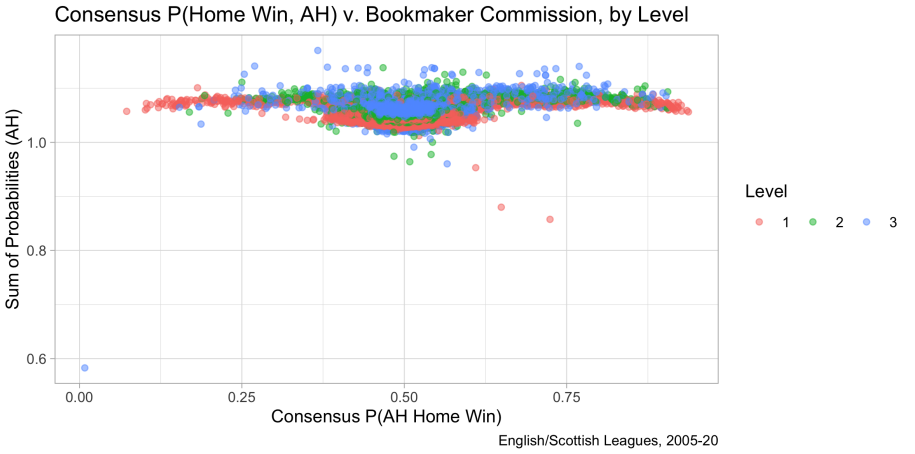

Figure 3.18: The sum of the $\mathbb{P}_{\text{cons}}$(AH Outcomes) v. $\mathbb{P}_{\text{cons}}$(AH Home Win), split by Level.

($\mathbb{P}_{\text{cons}} = 0.5$). In this market, the overround is much lower than for the 1X2 market, with η varying from just over 5% to just over 12%. We can notice more outliers too: these may be incorrectly copied odds rather than a genuine data-point: it is extremely unlikely the bookmakers have offered odds with a negative commission of around 40%. Our findings for the AH market are similar, with near identical distribution to the UO market.

These findings are backed up by the values in Table 3.6. In the 1X2 market, the mean overround in Level 1 is 7.1%, whereas in Level 3, this rises to 9.6%. In the UO market, it rises from 6.6% in Level 1 to 7.1% in Level 3; for the AH market, it is more stable, rising from 4.6% to 4.7%.

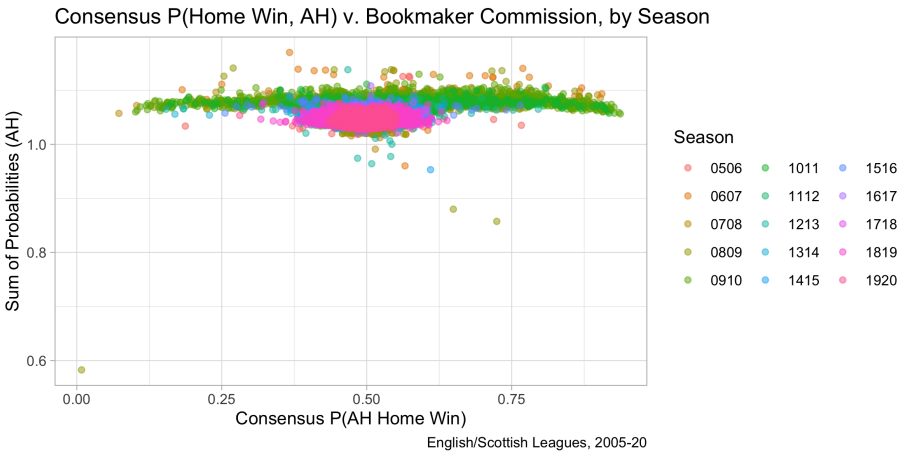

Figure 3.19: The sum of the $\mathbb{P}_{\text{cons}}$(AH Outcomes) v. $\mathbb{P}_{\text{cons}}$(AH Home Win), split by Season.

Table 3.6: The mean overroud $\bar{\eta}$ for different groups of our dataset.

	Market		
	1X2	Under/Over 2.5 Goals	Asian Handicap
Overall $\bar{\eta}$	.0825	.0684	.0470
Level 1 $\bar{\eta}$	.0707	.0662	.0461
Level 2 $\bar{\eta}$	.0838	.0685	.0473
Level 3 $\bar{\eta}$	.0961	.0712	.0477
Season 05/06 $\bar{\eta}$	.1110	.0798	.0436
Season 12/13 $\bar{\eta}$	.0769	.0635	.0400
Season 19/20 $\bar{\eta}$	.0659	.0592	.0468

Overround by Season

Importantly, we note the shape of these figures will be identical to those for overround by Level. Figures 3.15, 3.17, and 3.19 show the overround of each match, grouped by season, with the colour of the data points following the colours of a rainbow. Reds and yellows are for earlier seasons, greens and blues for 'middle' seasons, and violets and pinks for later (more-recent) seasons. In the 1X2 market 3.15, we notice the overround is decreasing over time, with the maximum overround decreasing from around 12% for the earliest seasons in our data to around 10%. The floor of the range hasn't moved from around 4%, though.

In the UO market, the variation in the overround has decreased massively: the pink datapoints for the 19/20 season cover a near-horizontal line at approximately 7%, with earlier seasons varying from approximately 0% to just over 10%.

The AH market overround hasn't changed much overtime, but the variation in the odds has: the consensus probabilities for 'middle' (green) varied from

around 0.12 to 0.8, whereas the pink datapoints are highly concentrated around 0.5: this indicates that as time is progressing, the optimal selection of the handicap is improving (optimal being the handicap resulting in 0.5 probability for both a Home and Away Win).

The values in Table 3.6 again back this up. Over time, the mean 1X2 overround has decreased from 11.1% to 6.6% and the mean UO overround from 8.0% to 5.9%. The mean AH overround has changed from 4.4% to 4.7%, despite being as low as 4.0% in 2012/13 (half-way through our dataset).

These values indicate bookmakers earn more commission in the ternary 1X2 market[4] than for the binary AH and UO markets.

[4]Three outcomes can be bet on.

3.5 Conclusion

In this section, we have found that across the English and Scottish football league pyramids, bookmakers are most accurate in the 1X2 market (excluding performance on draws), followed by the Asian Handicap market, with poor accuracy in the Under/Over 2.5 Goals market. Betting odds are, surprisingly, more accurate in Level 3—the lower leagues involving semi-professional sides—than in Level 2—wholly professional lower leagues; performance in Level 1 was greater than all other levels. Finally, in Section 3.4, we have shown the bookmaker commission is highest in the 1X2 market, highest in lower levels, and is reducing over time.

In Chapter 4, we apply our findings into the creation of a proposed method of placing bets in an attempt to use bookmaker accuracy to turn a profit.

Chapter 4

A proposed method of placing bets, using our findings.

In this chapter, we will use our findings from Chapters 2 and 3 to attempt to "beat the bookies" with their own odds. We will do this by accepting bookmaker accuracy is high, and so if we bet according to their favourites, we should be able to win the majority of our "bets" and thus make a profit.

4.1 The Method

We use the scheme in Equation 4.1 to choose our bet on match i market m, $B_{m,i}$, based on probability of the event $p_{m,i}$, where μ_m is the market mean and σ_m is the market standard deviation (these are figures calculated up to the match i).

$$B_{m,i} = \begin{cases} p_{m,i} < \mu_m + 0.5\sigma_m & B_{m,i} = 0 \\ \mu_m + 0.5\sigma_m \leq p_{m,i} < \mu_m + \sigma_m & B_{m,i} = 1 \\ \mu_m + \sigma_m \leq p_{m,i} < \mu_m + 1.5\sigma_m & B_{m,i} = 2 \\ \mu_m + 1.5\sigma_m \leq p_{m,i} & B_{m,i} = 3 \end{cases} \tag{4.1}$$

4.1.1 An Alternate Method Excluding Poor Performing Leagues

4.2 Results

4.3 Comparison

4.4 Conclusion

Chapter 5

Conclusion

In this final chapter, we will discuss what we have found throughout the dissertation, areas for future research, and, from a self-reflecting view, challenges I have been faced with throughout the duration of the project, and how I overcame them.

5.1 Our Findings

5.2 Challenges

Bibliography

Adèr, Herman J. (2008). "Phases and initial steps in data analysis". In: *Advising on Research Methods: A consultant's companion*. Chap. 14, pp. 333–357.

Ajadi, Theo et al. (2020). *Home truths: Annual Review of Football Finance 2020*. URL: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/sports-business-group/deloitte-uk-annual-review-of-football-finance-2020.pdf (visited on 03/25/2021).

Alto, Valentina (2019). *PCA: Eigenvectors and Eigenvalues*. URL: https://towardsdatascience.com/pca-eigenvectors-and-eigenvalues-1f968bc6777a (visited on 03/24/2021).

American Psychiatric Association (2018). "DSM5 Diagnostic Criteria: Gambling Disorder". In: *Diagnostic and Statistical Manual of Mental Disorders (5th Edition)*.

Angelini, Giovanni and Luca De Angelis (2019). "Efficiency of online football betting markets". In: *International Journal of Forecasting* 35.2, pp. 712–721.

Auguie, Baptiste (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. URL: https://CRAN.R-project.org/package=gridExtra.

Barrabi, Thomas (2020). *How does NFL's salary cap work?* URL: https://www.foxbusiness.com/sports/nfl-salary-cap-rules-explained (visited on 03/21/2021).

bet365 (n.d.). *Help: Soccer*. URL: https://help.bet365.com/product-help/sports/Rules/Soccer (visited on 01/12/2021).

Boscá, José E et al. (2009). "Increasing offensive or defensive efficiency? An analysis of Italian and Spanish football". In: *Omega* 37.1, pp. 63–78.

Buchdahl, Joseph (n.d.[a]). *Notes for Football Data*. URL: http://www.football-data.co.uk/notes.txt (visited on 10/02/2020).

— (n.d.[b]). *Pinnacle Betting Expert and Author: Joseph Buchdahl*. Pinnacle. URL: https://www.pinnacle.com/en/betting-resources/author/Joseph-Buchdahl (visited on 01/08/2021).

— (n.d.[c]). *What is Football-Data?* Football Data. URL: http://football-data.co.uk/#intro (visited on 01/08/2021).

Cain, Michael et al. (2000). "The favourite-longshot bias and market efficiency in UK football betting". In: *Scottish Journal of Political Economy* 47.1, pp. 25–36.

Cambridge Dictionary, ed. (n.d.). *COMPETITIVE BALANCE — definition in the Cambridge English Dictionary*. URL: https://dictionary.cambridge.org/us/dictionary/english/competitive-balance (visited on 01/12/2021).

Chen, James (2020). *Line Of Best Fit*. URL: https://www.investopedia.com/terms/l/line-of-best-fit.asp (visited on 04/06/2021).

Clapham, Christopher and James Nicholson (2014). *Oxford Concise Dictionary of Mathematics. 5th Edition*. Oxford University Press.

Conn, David (2017). *Premier League finances: the full club-by-club breakdown and verdict*. URL: https://www.theguardian.com/football/2017/jun/01/premier-league-finances-club-by-club (visited on 01/12/2021).

Constantinou, Anthony (2020). *Asian Handicap football betting with Rating-based Hybrid Bayesian Networks*. arXiv: 2003.09384 [stat.AP].

Cronin, Benjamin (2019). *American odds vs. Decimal odds*. Pinnacle. URL: https://www.pinnacle.com/en/betting-articles/educational/odds-formats-available-at-pinnacle-sports/ZWSJD9PPX69V3YXZ (visited on 01/08/2021).

Davey, Jacob (2020). *Why It's Time We Ban Betting Sponsors in English Football*. URL: https://versus.uk.com/2020/07/time-ban-betting-sponsors-english-football/ (visited on 02/15/2021).

DeepAI (n.d.). *Kernal Density Estimation Definition*. URL: https://deepai.org/machine-learning-glossary-and-terms/kernel-density-estimation (visited on 04/06/2021).

Draper, Norman R. and Harry Smith (1998). *Applied Regression Analysis. 3rd Edition*. John Wiley & Sons, New York.

English Football League (n.d.). *EFL Regulations, Section 3 - The League*. URL: https://www.efl.com/-more/governance/efl-rules--regulations/section-3---the-league/ (visited on 03/25/2021).

Fabrigar, Leandre R et al. (1999). "Evaluating the use of exploratory factor analysis in psychological research." In: *Psychological methods* 4.3, p. 272.

FIFA.com (2001). *FIFA Survey: approximately 250 million footballers worldwide*. URL: https://www.fifa.com/who-we-are/news/fifa-survey-approximately-250-million-footballers-worldwide-88048 (visited on 01/12/2021).

Fox, John and Sanford Weisberg (2019). *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Giulianotti, Richard (2012). "Football". In: *The Wiley-Blackwell Encyclopedia of Globalization*. American Cancer Society.

Goal (2019). *What is the MLS salary cap & how much are U.S. soccer players paid?* URL: https://www.goal.com/en-us/news/mls-salary-cap-how-much-us-soccer-players-paid/q015j4su3gb31bha41zto4fkb (visited on 03/21/2021).

Goossens, Kelly (2005). *Competitive Balance in European Football: Comparison by adapting measures: National Measure of Seasonal Imbalance and Top3*. University of Antwerp, Research Administration.

Grinstead, Charles Miller and James Laurie Snell (2012). *Introduction to probability*. American Mathematical Soc.

Hafez, Shamoon (2019). *Calciopoli: The scandal that rocked Italy and left Juventus in Serie B*. URL: https://www.bbc.co.uk/sport/football/49910626 (visited on 01/07/2021).

Henery, R. J. (1999). "Measures of Over-Round in Performance Index Betting". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 48.3,

pp. 435–439. ISSN: 00390526, 14679884. URL: http://www.jstor.org/stable/2681006.

Hoaglin, David C. (1977). "Mathematical Software and Exploratory Data Analysis". In: *Mathematical Software*. Ed. by John R. Rice. Academic Press, pp. 139–159. ISBN: 978-0-12-587260-7. DOI: https://doi.org/10.1016/B978-0-12-587260-7.50010-8.

Hoaglin, David C. and Roy E. Welsch (1978). "The Hat Matrix in Regression and ANOVA". In: *The American Statistician* 32.1, pp. 17–22.

Hyndman, Rob and Anne Koehler (2006). "Another look at measures of forecast accuracy". In: *International Journal of Forecasting* 22, pp. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.

Jolliffe, Ian T (1972). "Discarding variables in a principal component analysis. I: Artificial data". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 21.2, pp. 160–173.

Knuth, Kevin H. (2006). "Optimal data-based binning for histograms". In: *arXiv preprint physics/0605197*.

Lange, David (2020). *Teams of the German Bundesliga ranked by market (transfer) value of players in 2020*. URL: https://www.statista.com/statistics/283033/market-value-teams-german-football-bundesliga/#statisticContainer (visited on 01/12/2021).

Lorenz, M. O. (1905). "Methods of Measuring the Concentration of Wealth". In: *Publications of the American Statistical Association* 9.70, pp. 209–219.

MatterOfStats (n.d.). *What is Vig and Overround*. URL: http://www.matterofstats.com/what-is-vig-and-overround/ (visited on 04/06/2021).

Mendenhall, William, Robert J. Beaver, and Barbara M. Beaver (2013). *Introduction to Probability and Statistics (14th Edition)*. Brooks/Cole.

Meyer, David et al. (2020). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-4. URL: https://CRAN.R-project.org/package=e1071.

NBA.com (2020). *NBA Salary Cap set at $109.14 million for 2019-20*. URL: https://www.nba.com/news/nba-salary-cap-2019-20-season-set-10914-million (visited on 03/21/2021).

Owen, Alun (2009). "Dynamic bayesian forecasting models of football match outcomes". In: *2nd International Conference on Mathematics in Sport (IMA Sport 2009)* (Groningen, The Netherlands). Ed. by The Institute of Mathematics and its Applications (IMA).

Pearson, Karl (1901). "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, pp. 559–572.

Plotly.com (n.d.). *geom count in ggplot2*. URL: https://plotly.com/ggplot2/geom_count/ (visited on 03/28/2021).

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

RDocumentation (n.d.). *cut: Convert Numeric to Factor*. URL: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/cut (visited on 03/21/2021).

Rosen, Dan (2020). *NHL salary cap to remain same next season*. URL: https://www.nhl.com/news/nhl-salary-cap-to-remain-at-815-million/c-317372082 (visited on 03/21/2021).

Rottenberg, Simon (1956). "The Baseball Players' Labor Market". In: *Journal of Political Economy* 64.3, pp. 242–258. DOI: 10.1086/257790.

Score and Change, eds. (2020). *Overview of the 2019/2020 La Liga sponsors*. URL: https://www.scoreandchange.com/overview-2019-2020-la-liga-sponsors/ (visited on 02/15/2021).

Scottish Professional Football League (2021). *The Rules and Regulations of the Scottish Professional Football League*. URL: https://spfl.co.uk/admin/filemanager/images/shares/pdfs/SPFL%20Rules%20and%20Regulations%2016-Mar-21%20(MASTER%20COPY)%20CLEAN.pdf (visited on 03/25/2021).

Slowinski, Piper (2012). *Luxury Tax*. URL: https://library.fangraphs.com/business/luxury-tax/ (visited on 03/21/2021).

Smyth, Rob (2018). *World Cup stunning moments: West Germany 1-0 Austria in 1982*. URL: https://www.theguardian.com/football/blog/2014/feb/25/world-cup-25-stunning-moments-no3-germany-austria-1982-rob-smyth (visited on 01/07/2021).

Štrumbelj, Erik (2014). "On determining probability forecasts from betting odds". In: *International journal of forecasting* 30.4, pp. 934–943.

The Football Association Premier League Limited, ed. (2019). *Premier League Handbook, Season 2019/20*.

The World Bank (2018). *Gini index (World Bank estimate*. URL: https://data.worldbank.org/indicator/SI.POV.GINI?view=map (visited on 01/12/2021).

UEFA.com (2020). *UEFA Europa Conference League: all you need to know*. URL: https://www.uefa.com/uefaeuropaconferenceleague/news/0264-10fe90612aa3-37b2bc77f89e-n.d.--europa-conference-league-lowdown/ (visited on 01/12/2021).

— (n.d.[a]). *Country Coefficients*. URL: https://www.uefa.com/memberassociations/uefarankings/country/#/yr/2021 (visited on 01/05/2021).

— (n.d.[b]). *Member associations*. URL: https://www.uefa.com/insideuefa/member-associations/ (visited on 01/18/2021).

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: https://www.stats.ox.ac.uk/pub/MASS4/.

Weisstein, Eric W. (n.d.). *Reflection*. URL: https://mathworld.wolfram.com/Reflection.html (visited on 03/28/2021).

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Wickham, Hadley and Dana Seidel (2020). *scales: Scale Functions for Visualization*. R package version 1.1.1. URL: https://CRAN.R-project.org/package=scales.

Wold, Svante, Kim Esbensen, and Paul Geladi (1987). "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3, pp. 37–52.

Wu, Shaomin (2012). "Warranty data analysis: A review". In: *Quality and Reliability Engineering International* 28.8, pp. 795–805.

Appendices

Appendix A

Definitions

A.1 Mathematical and Statistical

Unless stated, these definitions have been taken from the Fifth Edition of the Oxford Concise Dictionary of Mathematics, published in 2014 (Clapham and Nicholson, 2014).

(i) Central Limit Theorem — "[T]he distribution of the mean of a sequence of random variables tends to a normal distribution as the number in the sequence increases."

(ii) Cleaning (Data) — Modification, removal, or replacement of "coarse" ("heaped, censored and missing") data (Wu, 2012).

(iii) Confidence Interval — "An interval, calculated from a sample, which contains the value of a certain population parameter with a specified probability."

(iv) Distributions — "[This is] concerned with the way in which the probability of its taking a certain value, or a value within a certain interval, varies. It may be given by the cumulative distribution function[,] its probability mass function [or] its probability density function."

(v) Fit — "[T]he degree of of correspondence between the observations and the model's predictions."

(vi) Kernel Density Estimation — The "process of finding an estimate probability density function of a random variable. The estimation attempts to infer characteristics of a population, based on a finite data set. The data smoothing problem often is used in signal processing and data science, as it is a powerful way to estimate probability density. In short, the technique allows one to create a smooth curve given a set of random data" (DeepAI, n.d.).

(vii) Leverage — The amount of influence each data point y_i can have on each fitted y-value, $\hat{y}_j$ (Hoaglin and Welsch, 1978).

(viii) Line of Best Fit — " A line through a scatter plot of data points that best expresses the relationship between those points" (Chen, 2020).

(ix) Mode — "For a continuous random variable, [a mode is] a value at which the probability density function has a local maximum."

(x) Noise — "[...]random error or variation in observations which is not explained by the model."

(xi) Normal Distribution — "The continuous probability distribution wth a probability density function f given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right)$$

denoted by $\mathcal{N}o(\mu,\ \sigma^2)$" where μ is the mean and σ^2 the variance.

(xii) Odds — "[...]expressed in the form $r : s$ corresponding in theory to a probability of $\frac{r}{r+s}$ of winning."

(xiii) Probability — "[...]is a measure of the possibility of the event occuring as the result of an experiment." Note that for an event A, the probability of its complement, $\mathbb{P}(A') = 1 - \mathbb{P}(A)$.

(xiv) Skewness — "The amount of asymmetry of a distribution... If the distribution has a long tail to the left[...] it is said to be skewed to the left and to have negative skewness."

(xv) Standard Deviation, σ — "The positive square root of the variance, a commonly used measure of the dispersion of observations in a sample."

(xvi) Standard Error, SE — "The standard deviation of an estimator of a population parameter."

(xvii) Variance, σ^2 — "[...] equal to $E[(X-\mu)^2]$," the expected value of the squared difference between an observation and the mean.

A.2 Gambling Terms

(i) Bettor — Someone who places a bet. Also *punter*, *gambler*.

(ii) Bookmaker — Organisation that accepts and pays off bets. Also *house*, *bookie*.

(iii) Calciopoli — A match fixing scandal which occurred in the 2004/05 and 2005/06 seasons, where Italian Serie A teams (AC Milan, Fiorentina, Juventus, Lazio and Reggina) "systematically influenced referees." (Hafez, 2019).

(iv) Competitive Balance — "The situation in which no one business of a group of competing businesses has an unfair advantage over the others." (Cambridge Dictionary, n.d.).

(v) Disgrace of Gijón — A football match between West Germany and Austria at the 1982 FIFA World Cup where a "mutually suitable scoreline" was played out, ensuring both sides progressed to the knockout rounds, leading to the final pair of matches of World Cup group stages being played simultaneously (Smyth, 2018).

(vi) Favourite-Longshot Bias — An anomaly in betting markets where the favourites win more often then the market probabilities imply; longshots less often (Cain et al., 2000).

(vii) Overround — "Bookmakers are business people and the prices they offer include a profit margin, which is sometimes referred to as the 'vig' or 'vigorish' in the prices they offer[...]. Overround is [...] defined as the sum of the reciprocals of all prices in a given market. Some definitions of overround subtract 1 from this sum" (MatterOfStats, n.d.). n.b. we choose to define overround with this subtraction.

(viii) Stake — The amount of money placed onto a bet by the bettor (the amount of money *at stake*); *the ante*.

A.3 Acronyms

Football Associations	
UEFA	the Union of European Football Associations (Europe)
FA	the Football Association (England)
DFB	Deutscher Fußball-bund (Germany)
FFF	Fédération Française de Football (France)
FIGC	Federazione Italiana Giuoco Calcio (Italy)
FPF	Federação Portuguesa de Futebol (Portugal)
RFEF	Real Federación Española de Fútbol (Spain)
SFA	the Scottish Football Association (Scotland)
Football Leagues	
EPL	English Premier League
SPL	Scottish Premier League
EFL	English Football League (Tiers 2 to 4)
SFL	Scottish Football League (Tiers 2 to 3)

Appendix B

Chapter 2 Code

```
#### (1) ANALYSING THE ACCURACY OF BETTING ON ELITE EUROPEAN LEAGUES
#Directory, Environment, Packages ----
setwd("~/Desktop/University/University Year 3/331MP/Data")
rm(list=ls())
#install.packages('car');install.packages('MASS');install.packages('ggplot2');
#install.packages('gridExtra')
library(car); library(MASS); library(ggplot2); library(gridExtra)

### IDA; Analysis of one season, one league [chosen at random] ----
#Reading the data:
fr_l1_1617 <-
  read.csv("https://www.football-data.co.uk/mmz4281/1617/F1.csv")
fr_l1_1617 <-
  fr_l1_1617[,c("Div", "Date", "HomeTeam", "AwayTeam", "FTHG", "FTAG",
                "FTR", "BbAvH", "BbAvD", "BbAvA")]
fr_l1_1617 <- na.omit(fr_l1_1617)

#Finding consensus probabilities (we need to Normalise the inverse odds)
fr_l1_1617$AvgHProbPN <- with(fr_l1_1617, round(1/BbAvH, 4))
fr_l1_1617$AvgDProbPN <- with(fr_l1_1617, round(1/BbAvD, 4))
fr_l1_1617$AvgAProbPN <- with(fr_l1_1617, round(1/BbAvA, 4))
fr_l1_1617$Overround <- with(fr_l1_1617,
                             (AvgHProbPN + AvgDProbPN + AvgAProbPN))
fr_l1_1617$AvgHProb <- with(fr_l1_1617, round(AvgHProbPN/Overround,4))
fr_l1_1617$AvgDProb <- with(fr_l1_1617, round(AvgDProbPN/Overround,4))
fr_l1_1617$AvgAProb <- with(fr_l1_1617, round(AvgAProbPN/Overround,4))

#Simple calculations
basic.fr.ida <- matrix(c(mean(fr_l1_1617$AvgHProb),
                         mean(fr_l1_1617$AvgDProb),
                         mean(fr_l1_1617$AvgAProb),
                         sd(fr_l1_1617$AvgHProb),
                         sd(fr_l1_1617$AvgDProb),
                         sd(fr_l1_1617$AvgAProb)),
                       ncol=3, nrow=2, byrow=T)
rownames(basic.fr.ida) <- c('mean', 'sd')
colnames(basic.fr.ida) <- c('h', 'd', 'a')
```

```
basic.fr.ida;round(prop.table(table(fr_l1_1617$FTR)), 4)

##Histograms
idaHomeHIST <- ggplot(fr_l1_1617, aes(AvgHProb)) +
  geom_histogram(binwidth=0.05, fill="blue") +
  coord_cartesian(xlim=c(0,1)) + theme_light() +
  labs(title="Home Win", x=NULL, y=NULL)
idaAwayHIST <- ggplot(fr_l1_1617, aes(AvgAProb)) +
  geom_histogram(binwidth=0.05, fill="coral") +
  coord_cartesian(xlim=c(0,1)) + theme_light() +
  labs(title="Away Win", x=NULL, y=NULL)
idaDrawHIST <- ggplot(fr_l1_1617, aes(AvgDProb)) +
  geom_histogram(binwidth=0.05, fill="green4") +
  coord_cartesian(xlim=c(0,1)) + theme_light() +
  labs(title="Draw", x=NULL, y=NULL)
ida.histogram <-
  grid.arrange(idaHomeHIST, idaDrawHIST, idaAwayHIST, nrow=3, ncol=1,
               left="Frequency", bottom="Consensus Probability",
               top="Histograms of the Consensus Probabilities for each
               outcome\nFrench Ligue 1, 2016-17 Season")

ggsave(path="./writeup/images", filename="elite_01_idahist.png",
       plot=ida.histogram, unit="cm", width=15, height=18)

### EDA ----
##Reading the data using a For Loop:
#Define which countries and seasons we need to read:
countries <- c("de", "en", "es", "fr", "it", "po")
co.we <- c("D1", "E0", "SP1", "F1", "I1", "P1")
#n.b. The Premier League's code is 0; other countries are 1.
seasons <- c("0506", "0607", "0708", "0809", "0910", "1011", "1112",
             "1213", "1314", "1415", "1516", "1617", "1718", "1819",
             "1920")
eliteTemp <- NULL; elite <- NULL
for (i in seasons){
  for (j in 1:6){
    eliteTemp <- read.csv(paste0(
      'https://www.football-data.co.uk/mmz4281/',i,'/',co.we[j],'.csv'),
      fileEncoding = 'latin1')
    eliteTemp$Country <- with(eliteTemp, countries[j])
    eliteTemp$Season <- with(eliteTemp, i)
    if (i=="1920"){
      eliteTemp$BbAvH<-eliteTemp$AvgH; eliteTemp$BbAvA<-eliteTemp$AvgA
      eliteTemp$BbAvD<-eliteTemp$AvgD
    }
    else{}
    eliteTemp <- eliteTemp[ ,c("Div", "Date", "HomeTeam", "AwayTeam",
                               "FTHG", "FTAG", "FTR", "BbAvH", "BbAvD",
                               "BbAvA", "Country", "Season")]
    elite <- rbind(elite, eliteTemp)
  }
}
elite <- na.omit(elite)

```

```
#Finding underlying probabilities:
#Pre-Normalised Probabilities
elite$AvgHProbPN <- with(elite, round(1/BbAvH, 4))
elite$AvgDProbPN <- with(elite, round(1/BbAvD, 4))
elite$AvgAProbPN <- with(elite, round(1/BbAvA, 4))
#To normalise them:
elite$overround<-with(elite, (AvgHProbPN + AvgDProbPN + AvgAProbPN))
elite$AvgHProb <-with(elite, round(AvgHProbPN/overround, 4))
elite$AvgDProb <-with(elite, round(AvgDProbPN/overround, 4))
elite$AvgAProb <-with(elite, round(AvgAProbPN/overround, 4))

#For later analysis, we need the Correct/Incorrect Probabilities:
N<-nrow(elite); N
elite$Correct<-with(elite, rep(0, N))
elite$Incorr1<-with(elite, rep(0, N))
elite$Incorr2<-with(elite, rep(0, N))

for (i in 1:N){
  if ((elite$FTR[i])=="A"){
    elite$Correct[i]<-(elite$Correct[i] + elite$AvgAProb[i])
    elite$Incorr1[i]<-(elite$Incorr1[i] + elite$AvgDProb[i])
    elite$Incorr2[i]<-(elite$Incorr2[i] + elite$AvgHProb[i])}
  else if ((elite$FTR[i])=="H"){
    elite$Correct[i]<-(elite$Correct[i] + elite$AvgHProb[i])
    elite$Incorr1[i]<-(elite$Incorr1[i] + elite$AvgDProb[i])
    elite$Incorr2[i]<-(elite$Incorr2[i] + elite$AvgAProb[i])}
  else {
    elite$Correct[i]<-(elite$Correct[i] + elite$AvgDProb[i])
    elite$Incorr1[i]<-(elite$Incorr1[i] + elite$AvgAProb[i])
    elite$Incorr2[i]<-(elite$Incorr2[i] + elite$AvgHProb[i])}
}
elite$logCorrect<-with(elite, rep(0,N))
for (j in 1:N){elite$logCorrect[j]<-log(elite$Correct[j], base=exp(1))}

## Simple calculations
basic.elite <- matrix(c(mean(elite$AvgHProb), mean(elite$AvgDProb),
                        mean(elite$AvgAProb), sd(elite$AvgHProb),
                        sd(elite$AvgDProb), sd(elite$AvgAProb)),
                      ncol=3, nrow=2, byrow=T)
rownames(basic.elite) <- c('mean', 'sd')
colnames(basic.elite) <- c('h', 'd', 'a')
basic.elite; round(prop.table(table(elite$FTR)), 4) #Observed
    probabilities

### VISUAL ANALYSIS ----
## Boxplots
bp.home <- ggplot(elite, aes(x=FTR, y=AvgHProb)) +
  geom_boxplot(outlier.size=0.75, outlier.alpha=0.7, color="blue") +
  theme_light() + stat_boxplot(coef=5) +
  labs(x="Actual Result", y="Consensus Probability of a Home Win",
       caption="") + coord_cartesian(ylim=c(0,1))
bp.draw <- ggplot(elite, aes(x=FTR, y=AvgDProb)) +
  geom_boxplot(outlier.size=0.75, outlier.alpha=0.7, color="green4") +
  theme_light() + stat_boxplot(coef=5) +
```

```
  labs(x="Actual Result", y="Consensus Probability of a Draw",
       caption="") + coord_cartesian(ylim=c(0,1))
bp.away <- ggplot(elite, aes(x=FTR, y=AvgAProb)) +
  geom_boxplot(outlier.size=0.75, outlier.alpha=0.7, color="coral") +
  theme_light() + stat_boxplot(coef=5) +
  labs(x="Actual Result", y="Consensus Probability of an Away Win",
       caption="Elite European Leagues, 2005-2020") +
  coord_cartesian(ylim=c(0,1))

eda.bp.all <-
  grid.arrange(bp.home, bp.draw, bp.away, nrow=1, ncol=3, top="Boxplot
  of the Cons Probabilites offered for each outcome\nv. Actual Result")
ggsave(path="./writeup/images", filename="elite_02_boxplot.png",
       plot=eda.bp.all, unit="cm", width=15, height=10)

##Density Plots
eda.wins.dens.all <- ggplot(elite, aes(x=AvgHProb)) +
  geom_density(color="blue") +
  geom_density(data=elite, mapping=aes(x=AvgAProb), color="coral",
               show.legend=T) + coord_cartesian(xlim=c(0,1)) +
  labs(title="Home and Away Wins", caption="Elite Leagues, 2005-2020",
       x="Consensus Probability", y="density") + theme_light()
eda.draw.dens.all <- ggplot(elite, aes(x=AvgDProb)) +
  geom_density(color="green4") + coord_cartesian(xlim=c(0,1)) +
  labs(title="Draws", caption="", x="Consensus Probability",
       y="density") + theme_light()

eda.density.all <-
  grid.arrange(eda.wins.dens.all, eda.draw.dens.all, nrow=1, ncol=2,
               left="", bottom="", top="Density Plots for Consensus
               Probability of Each Outcome")
ggsave(path="./writeup/images", filename="elite_02_edadensall.png",
       plot=eda.density.all, unit="cm", width=15, height=7)

View(elite[elite$AvgDProb > 0.6,]) #Extremely high P(Draw)
View(elite[elite$AvgDProb > 0.35,]) #Extremely high P(Draw)
#Splitting these by league:
#Home Wins
eda.home.dens <- ggplot(elite, aes(x=AvgHProb, color=Country)) +
  geom_density() + coord_cartesian(xlim=c(0,1)) +
  labs(title="Home Win", x=NULL, y=NULL) + theme_light() +
  geom_vline(aes(xintercept=mean(AvgHProb)),linetype="dashed",size=0.3)+
  guides(y="none") + theme(legend.position="none")
#Away Wins
eda.away.dens <- ggplot(elite, aes(x=AvgAProb, color=Country)) +
  geom_density() + coord_cartesian(xlim=c(0,1)) +
  labs(title="Away Win", x=NULL, y=NULL) + theme_light() +
  geom_vline(aes(xintercept=mean(AvgAProb)),linetype="dashed",size=0.3)+
  guides(y="none") + theme(legend.position="none")
#Draws
eda.draw.dens <- ggplot(elite, aes(x=AvgDProb, color=Country)) +
  geom_density() + coord_cartesian(xlim=c(0,0.8)) +
  labs(title="Draw", x=NULL, y=NULL) +
  geom_vline(aes(xintercept=mean(AvgDProb)),linetype="dashed",size=0.3)+
```

```
  guides(y="none") + theme_light() +
  scale_colour_discrete(labels = c("Germany","England","Spain","France",
                                   "Italy","Portugal"))

eda.density <-
  grid.arrange(eda.home.dens, eda.draw.dens, eda.away.dens, nrow=3,
               ncol=1, left="", bottom="Consensus Probability",
               top="Density Plots for the Cons. Probability for each
               outcome\nElite European Leagues, 2005-2020")
ggsave(path="./writeup/images", filename="elite_02_edadens.png",
       plot=eda.density, unit="cm", width=15, height=18)

##Tile Plot
#We will group 5+ goals together
elite$FTHG.Tile <- with(elite,rep(0,N))
elite$FTAG.Tile <- with(elite,rep(0,N))
for (k in 1:N){
  if ((elite$FTHG[k])>=5){elite$FTHG.Tile[k] <- 5}
  else{elite$FTHG.Tile[k] <- elite$FTHG[k]}}
for (k in 1:N){
  if ((elite$FTAG[k])>=5){elite$FTAG.Tile[k] <- 5}
  else{elite$FTAG.Tile[k] <- elite$FTAG[k]}}

elite.tile <- ggplot(elite, aes(y=FTAG.Tile, x=FTHG.Tile)) +
  geom_tile(aes(fill = Correct)) +
  scale_fill_distiller(palette = "Greens", direction = 1,
                       name="Correct\nProbability") + theme_light() +
  labs(title="Match Result v. Correct Consensus Probability",
       x="Home Goals Scored", y="Away Goals Scored",
       caption="Elite European Leagues, 2005-2020") +
  scale_y_discrete(limits=factor(c(1:4, "5+"))) +
  scale_x_discrete(limits=factor(c(1:4, "5+"))) +
  geom_abline(intercept=0, slope=1) +
  coord_cartesian(xlim=c(0,5), ylim=c(0,5))

ggsave(path="./writeup/images", filename="elite_05_tile.png",
       plot=elite.tile, unit="cm", width=15, height=15)

tpbinsizes.elite <- table(elite$FTAG.Tile, elite$FTHG.Tile) #Bin sizes

### CORRELATION ANALYSIS ----

## Binning the data:-
#Home Wins:-
elite$AvgHProb.cut <- cut(elite$AvgHProb, 124, include.lowest=T)
#First, we cut the data into 100 'bins'
levels(elite$AvgHProb.cut) <-
  tapply(elite$AvgHProb, elite$AvgHProb.cut, mean)
#Tapply finds the mean of the bin, rather than taking the midpoint
elite.observed.probabilites.TabH <-
  prop.table(table(elite$FTR, elite$AvgHProb.cut), 2)[c(1, 2, 3),]
#The c(1,2,3) will remove any extra (blank) rows
elite.observed.probabilites.H <- elite.observed.probabilites.TabH[3,]
#[n,]; if n = : 1 Away; 2 Draw; 3 Home (alphabetic)
```

```
253 elite.bookmaker.probabilites.H <-
254   as.numeric(names(elite.observed.probabilites.H))
255 #Away Wins:-
256 elite$AvgAProb.cut <- cut(elite$AvgAProb, 124, inclues.lowest=T)
257 levels(elite$AvgAProb.cut) <-
258   tapply(elite$AvgAProb, elite$AvgAProb.cut, mean)
259 elite.observed.probabilites.TabA <-
260   prop.table(table(elite$FTR, elite$AvgAProb.cut), 2)[c(1, 2, 3), ]
261 elite.observed.probabilites.A <- elite.observed.probabilites.TabA[1, ]
262 elite.bookmaker.probabilites.A <-
263   as.numeric(names(elite.observed.probabilites.A))
264 #Draws:-
265 elite$AvgDProb.cut <- cut(elite$AvgDProb, 124, inclues.lowest=T)
266 levels(elite$AvgDProb.cut) <-
267   tapply(elite$AvgDProb, elite$AvgDProb.cut, mean)
268 elite.observed.probabilites.TabD <-
269   prop.table(table(elite$FTR, elite$AvgDProb.cut), 2)[c(1, 2, 3), ]
270 elite.observed.probabilites.D <- elite.observed.probabilites.TabD[2, ]
271 elite.bookmaker.probabilites.D <-
272   as.numeric(names(elite.observed.probabilites.D))
273
274 #Finding R-Squared and RMSE:
275 elite.bookmaker.probabilities <-
276   c(elite.bookmaker.probabilites.H, elite.bookmaker.probabilites.D,
277     elite.bookmaker.probabilites.A)
278 elite.observed.probabilities <-
279   c(elite.observed.probabilites.H, elite.observed.probabilites.D,
280     elite.observed.probabilites.A)
281 #Home Wins
282 elite.lm.home <-
283   lm(elite.observed.probabilites.H ~ elite.bookmaker.probabilites.H)
        #Creates the linear model
284 round(summary(elite.lm.home)$r.squared, 5)
285 round(sqrt(mean(elite.lm.home$residuals^2)), 5)
286 #Away Wins
287 elite.lm.away <-
288   lm(elite.observed.probabilites.A ~ elite.bookmaker.probabilites.A)
289 round(summary(elite.lm.away)$r.squared, 5)
290 round(sqrt(mean(elite.lm.away$residuals^2)), 5)
291 #Draws
292 elite.lm.draw <-
293   lm(elite.observed.probabilites.D ~ elite.bookmaker.probabilites.D)
294 round(summary(elite.lm.draw)$r.squared, 5)
295 round(sqrt(mean(elite.lm.draw$residuals^2)),5)
296 #Overall
297 elite.linear.model <-
298   lm(elite.observed.probabilities ~ elite.bookmaker.probabilities)
299 elite.rsqu <- round(summary(elite.linear.model)$r.squared, 5)
300 elite.rmse <- round(sqrt(mean(elite.linear.model$residuals^2)), 5)
301
302 #Final Plot
303 elite.scatter <- ggplot(data=NULL,aes()) + geom_smooth() +
304   geom_jitter(aes(x=elite.bookmaker.probabilites.H,
305                   y=elite.observed.probabilites.H),
```

```
306               col="blue", size=0.75, show.legend=T) +
307   geom_smooth(aes(x=elite.bookmaker.probabilites.H,
308                   y=elite.observed.probabilites.H),
309               col="blue", method=lm, alpha=.15, size=0.5) +
310   geom_jitter(aes(x=elite.bookmaker.probabilites.D,
311                   y=elite.observed.probabilites.D),
312               col="green4", size=0.75, show.legend=T) +
313   geom_smooth(aes(x=elite.bookmaker.probabilites.D,
314                   y=elite.observed.probabilites.D),
315               col="green4", method=lm, alpha=.15, size=0.5) +
316   geom_jitter(aes(x=elite.bookmaker.probabilites.A,
317                   y=elite.observed.probabilites.A),
318               col="coral", size=0.75, show.legend = T) +
319   geom_smooth(aes(x=elite.bookmaker.probabilites.A,
320                   y=elite.observed.probabilites.A),
321               col="coral", method=lm, alpha=.15, size=0.5) +
322   geom_abline(intercept = 0, slope = 1, linetype="dashed") +
323   labs(title="Consensus v. Observed Probabilities",
324        x="Consensus Probability", y="Observed Probability",
325        caption="Elite Euro. Leagues, 2005-2020\nBin Size: 250 Games") +
326   coord_cartesian(xlim=c(0, 1), ylim=c(0, 1)) + theme_light()
327
328 ggsave(path = "./writeup/images", filename = "elite_03_scatter.png",
329        plot=elite.scatter, unit="cm", width=15, height=10)
330
331 ### HAT VALUES AND LEVERAGE PLOT ----
332 hats <- as.data.frame(hatvalues(elite.lm.draw))
333 leverageCrit <- (2*3)/nrow(hats)
334 hats[hats$`hatvalues(elite.lm.draw)` > leverageCrit,]
335 drawLevPlot <-
336   leveragePlot(elite.lm.draw, elite.bookmaker.probabilites.D,
        col="green4",
337                id = list(method=list(abs(residuals(elite.lm.draw,
338                                                    type="pearson"))),
339                          n=10, cex=0.6, col="red"),
340                xlab = "Consensus Probability (Draw) | Others",
341                ylab = "Observed Probability (Draw) | Others")
342 #png(path = "./writeup/images", filename = "elite_03_drawLeverage.png",
      plot=drawLevPlot, units="cm", width=20, height=10, res=450)
343
344 ### PREDICTIVE PERFORMANCE (Overall P1 and P2 Values) ----
345 #Calculating Owen (2009)'s P1 and P2 values for overall (all countries)
346 P1 <- exp( (1/N)*sum(elite$logCorrect) )
347 P2 <- (1/N)*sum( (1-elite$Correct)**2 +
348                  (elite$Incorr1)**2 + (elite$Incorr2)**2 )
349
350 ### MODELS FOR EACH LEAGUE AND COUNTRY ----
351 RSqu.H <- NULL; RSqu.D <- NULL; RSqu.A <- NULL; RSqu.O <- NULL
352 RMSE.H <- NULL; RMSE.D <- NULL; RMSE.A <- NULL; RMSE.O <- NULL
353 p1.split <- NULL; p2.split <- NULL
354
355 for(i in countries){
356   modelTempData <- elite[elite$Country==i, ]
357   #Bins
```

```
358  modelTempData$AvgHProb.cut <- cut(modelTempData$AvgHProb, 20,
359                                    include.lowest=T)
360  modelTempData$AvgDProb.cut <- cut(modelTempData$AvgDProb, 5,
361                                    include.lowest=T)
362  modelTempData$AvgAProb.cut <- cut(modelTempData$AvgAProb, 20,
363                                    include.lowest=T)
364  #Means of each bin
365  levels(modelTempData$AvgHProb.cut) <-
366    tapply(modelTempData$AvgHProb, modelTempData$AvgHProb.cut, mean)
367  levels(modelTempData$AvgDProb.cut) <-
368    tapply(modelTempData$AvgDProb, modelTempData$AvgDProb.cut, mean)
369  levels(modelTempData$AvgAProb.cut) <-
370    tapply(modelTempData$AvgAProb, modelTempData$AvgAProb.cut, mean)
371  #Observed Probability for each bin
372  modelTemp.obs.prob.tabH <-
373    prop.table(table(modelTempData$FTR,
374                     modelTempData$AvgHProb.cut), 2)[c(1, 2, 3), ]
375  modelTemp.obs.prob.tabD <-
376    prop.table(table(modelTempData$FTR,
377                     modelTempData$AvgDProb.cut), 2)[c(1, 2, 3), ]
378  modelTemp.obs.prob.tabA <-
379    prop.table(table(modelTempData$FTR,
380                     modelTempData$AvgAProb.cut), 2)[c(1, 2, 3), ]
381  modelTemp.obs.prob.H <- modelTemp.obs.prob.tabH[3, ]
382  modelTemp.obs.prob.D <- modelTemp.obs.prob.tabD[2, ]
383  modelTemp.obs.prob.A <- modelTemp.obs.prob.tabA[1, ]
384  #Finds the bookmaker probabilities for each group and creates vectors
385  modelTemp.boo.prob.H <- as.numeric(names(modelTemp.obs.prob.H))
386  modelTemp.boo.prob.D <- as.numeric(names(modelTemp.obs.prob.D))
387  modelTemp.boo.prob.A <- as.numeric(names(modelTemp.obs.prob.A))
388  modelTemp.bookmakers <- c(modelTemp.boo.prob.H, modelTemp.boo.prob.D,
389                            modelTemp.boo.prob.A)
390  modelTemp.observed <- c(modelTemp.obs.prob.H, modelTemp.obs.prob.D,
391                            modelTemp.obs.prob.A)
392  #Model creation
393  modelTempH <- lm(modelTemp.obs.prob.H~modelTemp.boo.prob.H)
394  modelTempD <- lm(modelTemp.obs.prob.D~modelTemp.boo.prob.D)
395  modelTempA <- lm(modelTemp.obs.prob.A~modelTemp.boo.prob.A)
396  modelTempO <- lm(modelTemp.observed~modelTemp.bookmakers)
397  #Finding values
398  RSqu.H <- c(RSqu.H, round(summary(modelTempH)$r.squared, 5))
399  RMSE.H <- c(RMSE.H, round(sqrt(mean(modelTempH$residuals^2)), 5))
400  RSqu.D <- c(RSqu.D, round(summary(modelTempD)$r.squared, 5))
401  RMSE.D <- c(RMSE.D, round(sqrt(mean(modelTempD$residuals^2)), 5))
402  RSqu.A <- c(RSqu.A, round(summary(modelTempA)$r.squared, 5))
403  RMSE.A <- c(RMSE.A, round(sqrt(mean(modelTempA$residuals^2)), 5))
404  RSqu.O <- c(RSqu.O, round(summary(modelTempO)$r.squared, 5))
405  RMSE.O <- c(RMSE.O, round(sqrt(mean(modelTempO$residuals^2)), 5))
406
407  p1.temp <- exp((1/(nrow(modelTempData)))*sum(modelTempData$logCorrect))
408  p2.temp <- (1/(nrow(modelTempData))) *
409    sum( (1-modelTempData$Correct)**2 + (modelTempData$Incorr1)**2 +
410           (modelTempData$Incorr2)**2 )
411
```

```
  p1.split <- c(p1.split, round(p1.temp, 5))
  p2.split <- c(p2.split, round(p2.temp, 5))
}
#Putting this into an easy-to-see Table:
league.values <- matrix(c(RSqu.H, RMSE.H, RSqu.D, RMSE.D, RSqu.A,
                          RMSE.A, RSqu.O, RMSE.O, p1.split, p2.split),
                        ncol=6, byrow=T)
rownames(league.values) <- c("RSqu.H", "RMSE.H", "RSqu.D", "RMSE.D",
                             "RSqu.A", "RMSE.A", "RSqu.O", "RMSE.O",
                             "p1.split", "p2.split")
colnames(league.values) <- countries
league.values <- as.table(league.values)

#For each season:
rsqu.season <- NULL; rmse.season <- NULL
p1.season <- NULL; p2.season <- NULL
for(i in seasons){
  modelTempData <- elite[elite$Season==i, ]
  modelTempData$AvgHProb.cut <-
    cut(modelTempData$AvgHProb, 10, include.lowest=T)
  modelTempData$AvgDProb.cut <-
    cut(modelTempData$AvgDProb, 5, include.lowest=T)
  modelTempData$AvgAProb.cut <-
    cut(modelTempData$AvgAProb, 10, include.lowest=T)
  #Finds the mean of each group (cut)
  levels(modelTempData$AvgHProb.cut) <-
    tapply(modelTempData$AvgHProb, modelTempData$AvgHProb.cut, mean)
  levels(modelTempData$AvgDProb.cut) <-
    tapply(modelTempData$AvgDProb, modelTempData$AvgDProb.cut, mean)
  levels(modelTempData$AvgAProb.cut) <-
    tapply(modelTempData$AvgAProb, modelTempData$AvgAProb.cut, mean)
  #Finds the observed probability for each cut
  modelTemp.obs.prob.tabH <-
    prop.table(table(modelTempData$FTR,
                     modelTempData$AvgHProb.cut), 2)[c(1, 2, 3), ]
  modelTemp.obs.prob.tabD <-
    prop.table(table(modelTempData$FTR,
                     modelTempData$AvgDProb.cut), 2)[c(1, 2, 3), ]
  modelTemp.obs.prob.tabA <-
    prop.table(table(modelTempData$FTR,
                     modelTempData$AvgAProb.cut), 2)[c(1, 2, 3), ]
  modelTemp.obs.prob.H <- modelTemp.obs.prob.tabH[3, ]
  modelTemp.obs.prob.D <- modelTemp.obs.prob.tabD[2, ]
  modelTemp.obs.prob.A <- modelTemp.obs.prob.tabA[1, ]
  #Finds the bookmaker probabilities for each group and creates vectors
  modelTemp.boo.prob.H <- as.numeric(names(modelTemp.obs.prob.H))
  modelTemp.boo.prob.D <- as.numeric(names(modelTemp.obs.prob.D))
  modelTemp.boo.prob.A <- as.numeric(names(modelTemp.obs.prob.A))
  modelTemp.bookmakers <-
    c(modelTemp.boo.prob.H, modelTemp.boo.prob.D, modelTemp.boo.prob.A)
  modelTemp.observed <-
    c(modelTemp.obs.prob.H, modelTemp.obs.prob.D, modelTemp.obs.prob.A)

  #Making the model
```

```
  modelTemp0 <- lm(modelTemp.observed~modelTemp.bookmakers)
  #Finding values
  rsqu.season <- c(rsqu.season, round(summary(modelTemp0)$r.squared, 5))
  rmse.season <- c(rmse.season,
                   round(sqrt(mean(modelTemp0$residuals^2)), 5))

  p1.temp <- exp((1/(nrow(modelTempData)))*sum(modelTempData$logCorrect))
  p2.temp <- (1/(nrow(modelTempData))) *
    sum( (1-modelTempData$Correct)**2 + (modelTempData$Incorr1)**2 +
          (modelTempData$Incorr2)**2 )

  p1.season <- c(p1.season, round(p1.temp, 5))
  p2.season <- c(p2.season, round(p2.temp, 5))
}

season.values <- matrix(c(rsqu.season, rmse.season, p1.season,
                          p2.season), ncol=15, byrow=T)
rownames(season.values) <- c("rsqu", "rmse", "p1", "p2")
colnames(season.values) <- seasons

#Plotting Season Values
rsqu.se.plot <- ggplot(NULL, aes(y=rsqu.season, x=c(2005:2019))) +
  geom_jitter(color="violetred1")+theme_light() + labs(x = 'Year', y =
      'R2')+
  geom_smooth(method = 'lm', color = 'violetred4', se = F)
rmse.se.plot <- ggplot(NULL, aes(y=rmse.season, x=c(2005:2019))) +
  geom_jitter(color="steelblue4")+theme_light() + labs(x = 'Year', y =
      'RMSE')+
  geom_smooth(method = 'lm', color = 'steelblue1', se = F)
p1.se.plot <- ggplot(NULL, aes(y=p1.season, x=c(2005:2019))) +
  geom_jitter(color="darkorange4")+theme_light() + labs(x = 'Year', y =
      'P1')+
  geom_smooth(method = 'lm', color = 'darkorange1', se = F)
p2.se.plot <- ggplot(NULL, aes(y=p2.season, x=c(2005:2019))) +
  geom_jitter(color="slateblue4")+theme_light() + labs(x = 'Year', y =
      'P2')+
  geom_smooth(method = 'lm', color = 'slateblue1', se = F)

seasontimeplot <-
  grid.arrange(rsqu.se.plot, rmse.se.plot, p1.se.plot, p2.se.plot,
               nrow = 2, top = 'Elite Leagues Accuracy Statistics over
                   Time')

ggsave(path="./writeup/images", filename="elite_06_seasontimeplot.png",
       plot=seasontimeplot, unit="cm", width=20, height=15)

### COMPETITIVE BALANCE PCA ----

##LEAGUE MODEL:
#(For leagues, we have the comp. bal. statistics, unlike for seasons)
#We first define the statistics (Gini, NAMSI and K) from Goossens (05):
namsi <- c(0.374, 0.372, 0.364, 0.342, 0.418, 0.505)
kappa <- c(5.71, 5.79, 5.07, 6.00, 5.36, 4.07); invkap <- 1/kappa
gini  <- c(0.723, 0.826, 0.861, 0.784, 0.737, 0.898)
```

```
#We define IMBALANCE (Scale (standardise) each statistic above):
namsisc <- scale(namsi); invkapsc <- scale(invkap); ginisc <- scale(gini)
imbalance <- (namsisc[c(1:6),] + invkapsc[c(1:6),] + ginisc[c(1:6),])/3
#The [c(1:6),] cuts off the sd and mean attributes from the scaled data

#We define the LEVEL OF ATTACK - Shots Per Game / Goals Per Game.
attack <- NULL; attackPO <- NULL
for (l in 1:5){
  for (s in seasons){
    dataTemp <- read.csv(paste0(
      "https://www.football-data.co.uk/mmz4281/",s,"/",co.we[l],".csv"))
    dataTemp <- dataTemp[ ,c("FTHG", "FTAG", "HS", "AS")]
    dataTemp$totalGoals <- with(dataTemp, FTHG+FTAG)
    dataTemp$totalShots <- with(dataTemp, HS+AS)
    dataTemp<-na.omit(dataTemp)
    attack <- c(attack, mean(dataTemp$totalShots)/
                mean(dataTemp$totalGoals))
  }
}
for (s in seasons[13:15]){
  #Data for Po is only avaliable for the 17/18 season onwards.
  dataTemp <- read.csv(paste0(
    "https://www.football-data.co.uk/mmz4281/",s,"/", co.we[6], ".csv"))
  dataTemp <- dataTemp[ ,c("FTHG", "FTAG", "HS", "AS")]
  dataTemp$totalGoals <- with(dataTemp, FTHG+FTAG)
  dataTemp$totalShots <- with(dataTemp, HS+AS)
  attackPO <- c(attackPO, mean(dataTemp$totalShots)/
                mean(dataTemp$totalGoals))
}
attack <- matrix(attack, ncol=15, byrow = T)
attackPO <- matrix(attackPO, nrow=1, byrow = T)
colnames(attack) <- seasons; rownames(attack) <- countries[1:5]
colnames(attackPO) <- seasons[13:15]; rownames(attackPO) <- countries[6]

attack.league <-
  c(mean(attack[1,]), mean(attack[2,]), mean(attack[3,]),
    mean(attack[4,]), mean(attack[5,]), mean(attackPO[1,]))
attack.season <- NULL
for (i in 1:12){attack.season <- c(attack.season, mean(attack[,i]))}
for (i in 1:3){
  attack.season <- c(attack.season,
                     mean(c(attack[1,(i+12)], attack[2,(i+12)],
                            attack[3,(i+12)], attack[4,(i+12)],
                            attack[5,(i+12)], attackPO[1,i] )))
}

#We define PredAcc, as the normalised sum of the 4 predictive statistics
#We will take the inverse of RMSE and P2
#A high PredAcc value => better bookmaker performance
invrmse.l <- 1/RMSE.O; invp2.l <- 1/p2.split
rsqu.l.sc <- scale(RSqu.O); invrmse.l.sc <- scale(invrmse.l)
p1.l.sc <- scale(p1.split); invp2.l.sc <- scale(invp2.l)
predacc <- (rsqu.l.sc + invrmse.l.sc + p1.l.sc + invp2.l.sc)/4

```

```
pc.league <- matrix(c(imbalance, attack.league, predacc), ncol=3,
                    byrow=F)
colnames(pc.league) <- c("imbalance", "attack", "predacc")
rownames(pc.league) <- countries

league.model <- prcomp(pc.league)
league.model$rotation; summary(league.model)

##SEASON MODEL:
pc.season <- matrix(c(rsqu.season, (1/rmse.season), p1.season,
                      (1/p2.season), attack.season), ncol = 5, byrow=F)
colnames(pc.season) <- c("rsqu", "inv rmse", "p1", "inv p2", "attack")
rownames(pc.season) <- seasons
pc.season.sc <- scale(pc.season)

season.model <- prcomp(pc.season.sc)
summary(season.model); round(season.model$rotation,3)

##PLOTS:
#League Model Plots
leascree <- ggplot(NULL, aes(x = c(1:3), y = (league.model$sdev)^2)) +
  geom_line() + geom_point(size = 2) + theme_light() +
  geom_abline(slope = 0, intercept = 1, color = "red") +
  labs(x = "Principal Component", y = "Variances",
       title = "Screeplot of League PCA Components")

leacomps <- ggplot(NULL, aes(x = league.model$x[,1], y =
    league.model$x[,2],
                             label = countries)) + geom_jitter() +
  labs(x = 'Component 1', y = 'Component 2', title = "PC1 v. PC2") +
  theme_light() +
  geom_text(aes(x = league.model$x[,1], y = league.model$x[,2]-0.1))

leaguepca <- grid.arrange(leascree, leacomps, ncol = 2, top = "By-League
    PCA")

ggsave(path="./writeup/images", filename="elite_07a_leaguepca.png",
       plot=leaguepca, unit="cm", width=20, height=10)

#Season Model Plots
seascree <- ggplot(NULL, aes(x = c(1:5), y = (season.model$sdev)^2)) +
  geom_line() + geom_point(size = 2) + theme_light() +
  geom_abline(slope = 0, intercept = 1, color = "red") +
  labs(x = "Principal Component", y = "Variances",
       title = "Screeplot of Season PCA Components")

seacomps <- ggplot(NULL, aes(x = season.model$x[,1], y =
    season.model$x[,2],
                               label = seasons)) + geom_jitter() +
  labs(x = 'Component 1', y = 'Component 2', main = "PC1 v. PC2") +
  theme_light() +
  geom_text(aes(x = season.model$x[,1], y = season.model$x[,2]-0.2))

```

```
seasonpca <- grid.arrange(seascree, seacomps, ncol = 2, top = "By-Season
    PCA")

ggsave(path="./writeup/images", filename="elite_07b_seasonpca.png",
       plot=seasonpca, unit="cm", width=20, height=10)
```

Appendix C

Chapter 3 Code

```
#### (2) ANALYSING THE ACCURACY OF BETTING ON ENG/SCO FOOTBALL LEAGUES
#----
#N.B.:
#assume the working directory and libraries are set as in eliteleagues.R
#else run:
#setwd("~/Desktop/University/University Year 3/331MP/Data");rm(list=ls())
library(car); library(MASS); library(ggplot2); library(gridExtra)
### READING DATA ----
#We downloading data straight from football-data.co.uk

divisions <- c("E0", "E1", "E2", "E3", "EC", "SC0", "SC1", "SC2", "SC3")
levels <- c("1", "2", "3")
seasons <- c("0506", "0607", "0708", "0809", "0910", "1011", "1112",
             "1213", "1314", "1415", "1516", "1617", "1718", "1819",
             "1920")

enscoTemp <- NULL; ensco <- NULL
for (i in seasons){
  for (j in divisions){
    enscoTemp <- read.csv(paste0(
      "https://www.football-data.co.uk/mmz4281/",i,"/",j,".csv"),
      fileEncoding="latin1")
    #The above line will download and read the .csv file in one go.
    enscoTemp$Season <- with(enscoTemp, i)
    enscoTemp$Div <- with(enscoTemp, j)
    if (i=="1920"){
      enscoTemp$BbAvH <- enscoTemp$AvgH
      enscoTemp$BbAvA <- enscoTemp$AvgA
      enscoTemp$BbAvD <- enscoTemp$AvgD
      enscoTemp$BbAv.2.5 <- enscoTemp$Avg.2.5
      enscoTemp$BbAv.2.5.1 <- enscoTemp$Avg.2.5.1
      enscoTemp$BbAvAHH <- enscoTemp$AvgAHH
      enscoTemp$BbAvAHA <- enscoTemp$AvgAHA
      enscoTemp$BbAHh <- enscoTemp$AHh}
    else{}
    enscoTemp$Over2.5Odds <- enscoTemp$BbAv.2.5
    enscoTemp$Under2.5Odds <- enscoTemp$BbAv.2.5.1
```

```
    #Greater Than or Less Than don't copy through:
    #manual check confirms this is the right way round
    enscoTemp$HomeHandicap <- enscoTemp$BbAHh
    enscoTemp <- enscoTemp[,c("Div", "Date", "HomeTeam", "AwayTeam",
                              "FTHG", "FTAG", "FTR", "BbAvH", "BbAvD",
                              "BbAvA", "Over2.5Odds", "Under2.5Odds",
                              "HomeHandicap", "BbAvAHH", "BbAvAHA",
                              "Season")]
    ensco <- rbind(ensco, enscoTemp)
  }
}
ensco<-na.omit(ensco)

### NORMALISING PROBS, FINDING WINNING BETS ----

#Defining the league 'level': 1 - Elite (EPL, SPL, Championship);
#                             2 - Fully Professional Lower Leagues;
#                             3 - Semi-Professional Lower Leagues.
ensco$Level<-with(ensco, rep(0, nrow(ensco)))
for (k in 1:(nrow(ensco))){
  if (ensco$Div[k]=="E0"){ensco$Level[k]<-1}
  else if (ensco$Div[k]=="E1"){ensco$Level[k]<-1}
  else if (ensco$Div[k]=="E2"){ensco$Level[k]<-2}
  else if (ensco$Div[k]=="E3"){ensco$Level[k]<-2}
  else if (ensco$Div[k]=="EC"){ensco$Level[k]<-3}
  else if (ensco$Div[k]=="SC0"){ensco$Level[k]<-1}
  else if (ensco$Div[k]=="SC1"){ensco$Level[k]<-2}
  else if (ensco$Div[k]=="SC2"){ensco$Level[k]<-3}
  else if (ensco$Div[k]=="SC3"){ensco$Level[k]<-3}
  else{}
}

#Adding and normalising Probability columns
#Pre Normalised:-
  #1X2 Market
    ensco$AvgHProbPN <- with(ensco, round(1/BbAvH, 4))
    ensco$AvgDProbPN <- with(ensco, round(1/BbAvD, 4))
    ensco$AvgAProbPN <- with(ensco, round(1/BbAvA, 4))
  #Under/Over 2.5 goals market
    ensco$Over2.5ProbPN <- with(ensco, round(1/Over2.5Odds, 4))
    ensco$Under2.5ProbPN <- with(ensco, round(1/Under2.5Odds, 4))
  #Asian Handicap Markets
    ensco$AH.HProbPN <- with(ensco, round(1/BbAvAHH, 4))
    ensco$AH.AProbPN <- with(ensco, round(1/BbAvAHA, 4))
#Finding Overrounds:-
    ensco$OneXTwoOverround <-
      with(ensco, (AvgHProbPN + AvgDProbPN + AvgAProbPN))
    ensco$UnderOverOverround <-
      with(ensco, (Over2.5ProbPN + Under2.5ProbPN))
    ensco$AHOverround <- with(ensco, (AH.HProbPN + AH.AProbPN))
#Normalising
    ensco$AvgHProb <- with(ensco, round(AvgHProbPN/OneXTwoOverround, 4))
    ensco$AvgDProb <- with(ensco, round(AvgDProbPN/OneXTwoOverround, 4))
    ensco$AvgAProb <- with(ensco, round(AvgAProbPN/OneXTwoOverround, 4))
```

```
92      ensco$Over2.5Prob <-
93        with(ensco, round(Over2.5ProbPN/UnderOverOverround, 4))
94      ensco$Under2.5Prob <-
95        with(ensco, round(Under2.5ProbPN/UnderOverOverround, 4))
96      ensco$AH.HProb <- with(ensco, round(AH.HProbPN/AHOverround, 4))
97      ensco$AH.AProb <- with(ensco, round(AH.AProbPN/AHOverround, 4))
98
99  #A few important notes:
100 #Rangers had a -275 goal handicap v. East Fife; assume this meant -2.75:
101 ensco$HomeHandicap[ensco$HomeTeam=="Rangers" &
102                    ensco$Date=="11/01/14"] <- -2.75
103 #Hamilton had a 12.5 goal handicap v. Rangers; assume this meant 1.25:
104 ensco$HomeHandicap[ensco$HomeTeam=="Hamilton" &
105                    ensco$Date=="25/10/08"] <- 1.25
106
107 #"Correct" Results -- Straight forward for the 1X2 and U/O Markets:
108 N = nrow(ensco)
109 ensco$Correct1X2 <- with(ensco, rep(0,N))
110 ensco$IncorrectA1X2 <- with(ensco, rep(0,N))
111 ensco$IncorrectB1X2 <- with(ensco, rep(0,N))
112 ensco$TotGoals <- with(ensco, FTHG + FTAG)
113 ensco$CorrectUO <- with(ensco, rep(0,N))
114 ensco$IncorrectUO <- with(ensco, rep(0,N))
115
116 for (l in 1:N){
117   if (ensco$FTR[l] == "H"){
118     ensco$Correct1X2[l] <- ensco$Correct1X2[l] + ensco$AvgHProb[l]
119     ensco$IncorrectA1X2[l] <- ensco$IncorrectA1X2[l] + ensco$AvgAProb[l]
120     ensco$IncorrectB1X2[l] <- ensco$IncorrectB1X2[l] + ensco$AvgDProb[l]}
121   else if (ensco$FTR[l] == "D"){
122     ensco$Correct1X2[l] <- ensco$Correct1X2[l] + ensco$AvgDProb[l]
123     ensco$IncorrectA1X2[l] <- ensco$IncorrectA1X2[l] + ensco$AvgAProb[l]
124     ensco$IncorrectB1X2[l] <- ensco$IncorrectB1X2[l] + ensco$AvgHProb[l]}
125   else if (ensco$FTR[l] == "A"){
126     ensco$Correct1X2[l] <- ensco$Correct1X2[l] + ensco$AvgAProb[l]
127     ensco$IncorrectA1X2[l] <- ensco$IncorrectA1X2[l] + ensco$AvgHProb[l]
128     ensco$IncorrectB1X2[l] <- ensco$IncorrectB1X2[l] + ensco$AvgDProb[l]}
129   else{}
130   if (ensco$TotGoals[l] > 2.5){
131     ensco$CorrectUO[l] <- ensco$CorrectUO[l] + ensco$Over2.5Prob[l]
132     ensco$IncorrectUO[l] <- ensco$IncorrectUO[l] + ensco$Under2.5Prob[l]}
133   else if (ensco$TotGoals[l] < 2.5){
134     ensco$CorrectUO[l] <- ensco$CorrectUO[l] + ensco$Under2.5Prob[l]
135     ensco$IncorrectUO[l] <- ensco$IncorrectUO[l] + ensco$Over2.5Prob[l]}
136   else{}
137 }
138
139 ensco$uo.res <- NULL
140 for (b in 1:N){
141   if (ensco$TotGoals[b] > 2.5){ensco$uo.res[b] <- "over"}
142   else {ensco$uo.res[b] <- "under"}
143 }
144
145 #For Asian Handicaps, we need to work out the winner(s)
```

```
ensco$FTHG.ah <- with(ensco, rep(0,N))
for (m in 1:N){ensco$FTHG.ah[m] <- ensco$FTHG[m] + ensco$HomeHandicap[m]}

#There are 3 types of AH bets: Assume the bookie bets on the HOME team
#   Integer: e.g. +1 Handicap for home team
         ##  EVENT                                #Winner
         # - If there is a draw, or home team wins Home
         # - If away teams by 1 goal             Void (stake refund)
         # - If away team wins by more than 1 goal Away
#   Half: e.g. +0.5 Handicap for the home team
         # - If the home team wins, or it is a draw Home wins
         # - If the away team wins               Away
#   Quarter: e.g. +0.75 Handicap for the home team
#        Half the stake goes to +1, Half goes to +0.5
         # - If the home team wins               Home
         # - If the game is a draw               Home wins
         # - If away team wins by 1 goal         HalfAway wins
         # - If away wins by more than 1 goal    Away wins

ensco$ah.gap <- with(ensco, FTHG.ah - FTAG); ensco$ah.res <- NULL
for (n in 1:N){
  if (ensco$ah.gap[n]<(-0.25)){ensco$ah.res[n]<-"aw"}
  else if (ensco$ah.gap[n]==(-0.25)){ensco$ah.res[n]<-"hfaw"}
  else if (ensco$ah.gap[n]==0){ensco$ah.res[n]<-"vo"}
  else if (ensco$ah.gap[n]==0.25){ensco$ah.res[n]<-"hfhm"}
  else if (ensco$ah.gap[n]>0.25){ensco$ah.res[n]<-"hm"}
  else{}
}

### BASIC CALCULATIONS ----
#1X2:
basic.1x2 <- NULL
for (a in 1:3){basic.1x2 <- c(basic.1x2,
                              mean(ensco$AvgHProb[ensco$Level == a]))}
for (a in 1:3){basic.1x2 <- c(basic.1x2,
                              mean(ensco$AvgDProb[ensco$Level == a]))}
for (a in 1:3){basic.1x2 <- c(basic.1x2,
                              mean(ensco$AvgAProb[ensco$Level == a]))}
for (a in 1:3){basic.1x2 <- c(basic.1x2,
                              sd(ensco$AvgHProb[ensco$Level == a]))}
for (a in 1:3){basic.1x2 <- c(basic.1x2,
                              sd(ensco$AvgDProb[ensco$Level == a]))}
for (a in 1:3){basic.1x2 <- c(basic.1x2,
                              sd(ensco$AvgAProb[ensco$Level == a]))}
basic.1x2 <- matrix(c(basic.1x2), ncol=3, byrow=T)
colnames(basic.1x2) <- 1:3
rownames(basic.1x2) <- c("1x2 Home Mean", "1x2 Draw Mean",
                         "1x2 Away Mean", "1x2 Home SD", "1x2 Draw SD",
                         "1x2 Away SD")
#Under/Over:
basic.uo <- NULL
for (a in 1:3){basic.uo <- c(basic.uo,
                             mean(ensco$Under2.5Prob[ensco$Level == a]))}
for (a in 1:3){basic.uo <- c(basic.uo,
```

```
                          mean(ensco$Over2.5Prob[ensco$Level == a]))}
for (a in 1:3){basic.uo <- c(basic.uo,
                          sd(ensco$Under2.5Prob[ensco$Level == a]))}
for (a in 1:3){basic.uo <- c(basic.uo,
                          sd(ensco$Over2.5Prob[ensco$Level == a]))}
basic.uo <- matrix(c(basic.uo), ncol=3, byrow=T)
colnames(basic.uo) <- 1:3
rownames(basic.uo) <- c("Under 2.5 Mean", "Over 2.5 Mean",
                        "Under 2.5 SD", "Over 2.5 SD")
#Asian Handicaps:
basic.ah <- NULL
for (a in 1:3){basic.ah <- c(basic.ah,
                          mean(ensco$AH.HProb[ensco$Level == a]))}
for (a in 1:3){basic.ah <- c(basic.ah,
                          mean(ensco$AH.AProb[ensco$Level == a]))}
for (a in 1:3){basic.ah <- c(basic.ah,
                          sd(ensco$AH.HProb[ensco$Level == a]))}
for (a in 1:3){basic.ah <- c(basic.ah,
                          sd(ensco$AH.AProb[ensco$Level == a]))}
basic.ah <- matrix(c(basic.ah), ncol=3, byrow=T)
colnames(basic.ah) <- 1:3
rownames(basic.ah) <- c("AH Home Mean", "AH Away Mean",
                        "AH Home SD", "AH Away SD")

basic.calcs <- round(rbind(basic.1x2, basic.uo, basic.ah),4)

#Observed Probabilities
obsprob.1x2tab <- round(prop.table(table(ensco$FTR, ensco$Level),2),
                        4)[c(1,2,3), c(1,2,3)]
obsprob.uotab <- round(prop.table(table(ensco$uo.res, ensco$Level),2),
                       4)[c(1,2), c(1,2,3)]
obsprob.ahtab <- round(prop.table(table(ensco$ah.res, ensco$Level),2),
                       4)[c("hm", "hfhm", "vo", "hfaw", "aw"),]

#To find the % of AH full wins that are home wins
print(nrow(ensco[ensco$ah.res == "hm",]) /
  (nrow(ensco[ensco$ah.res == "hm",]) + nrow(ensco[ensco$ah.res ==
      "aw",])))

#Finding this by Level
AH.Basic.Proportions <- NULL
for (i in levels){
  tempH <- nrow(ensco[ensco$ah.res == "hm" & ensco$Level == i,])
  tempA <- nrow(ensco[ensco$ah.res == "aw" & ensco$Level == i,])
  tempProp <- tempH / (tempH + tempA)
  AH.Basic.Proportions <- c(AH.Basic.Proportions, tempProp)
}
AH.Basic.Proportions <- as.matrix(AH.Basic.Proportions, nrow = 1, ncol =
    3)
rownames(AH.Basic.Proportions) <- paste("Level",levels)
AH.Basic.Proportions

basic.calcs; obsprob.1x2tab; obsprob.uotab; obsprob.ahtab

```

```
### PLOTS AND VISUAL ANALYSIS ----
#Density Plots
dens1x2ha <- ggplot(ensco, aes(x=AvgHProb, color="Home Win")) +
  geom_density() +
  geom_density(data=ensco, mapping=aes(x=AvgAProb, color="Away Win")) +
  scale_color_manual(name="Bet Type",
                     values=c("Home Win" = "blue",
                              "Away Win" = "coral")) +
  labs(x="Consensus Probabilitiy", y="Density",
       caption="English and Scottish Leagues, 2005-2020",
       title="Home and Away Wins in\nthe 1X2 Market") +
  theme_light() + coord_cartesian(xlim=c(0,1))

dens1x2d <- ggplot(ensco, aes(x=AvgDProb, color="Draw")) +
  geom_density() +
  labs(x="Consensus Probability", y="Density",
       caption="English and Scottish Leagues, 2005-2020",
       title="\nDraws in \nthe 1X2 Market") +
  scale_color_manual(name="Bet Type", values=c("Draw" = "green4")) +
  theme_light() + coord_cartesian(xlim=c(0,1))

densuo <- ggplot(ensco, aes(x=Under2.5Prob, color="Under")) +
  geom_density() +
  geom_density(data=ensco, aes(x=Over2.5Prob, color="Over")) +
  labs(x="Consensus Probability", y="Density",
       caption="English and Scottish Leagues, 2005-2020",
       title="Under/Over 2.5\nGoals Market") +
  scale_color_manual(name="Bet Type",
                     values=c("Under" = "red", "Over" = "green")) +
  theme_light() + coord_cartesian(xlim=c(0,1))

densah <- ggplot(ensco, aes(x=AH.HProb, color="AH Home")) +
  geom_density()+
  geom_density(data=ensco, aes(x=AH.AProb, color="AH Away")) +
  labs(x="Consensus Probability", y="Density",
       caption="English and Scottish Leagues, 2005-2020",
       title="Home and Away Wins in the\nAsian Handicap Market") +
  scale_color_manual(name="Bet Type", values=c("AH Home" = "blue",
                                               "AH Away" = "coral")) +
  theme_light() + coord_cartesian(xlim=c(0,1))

# Saving the plots:
ggsave(path = "./writeup/images", filename = "ensco_01_dens1x2ha.png",
       plot=dens1x2ha, unit="cm", width=15, height=10)
ggsave(path = "./writeup/images", filename = "ensco_02_dens1x2d.png",
       plot=dens1x2d, unit="cm", width=15, height=10)
ggsave(path = "./writeup/images", filename = "ensco_03_densuo.png",
       plot=densuo, unit="cm", width=15, height=10)
ggsave(path = "./writeup/images", filename = "ensco_04_densah.png",
       plot=densah, unit="cm", width=15, height=10)

dens <- grid.arrange(dens1x2ha, dens1x2d, densuo, densah, ncol=2, nrow=2,
                     top = "Density Plots of Consensus Probabilities")
ggsave(path = "./writeup/images", filename = "ensco_04A_densities.png",
```

```
306       plot=dens, unit="cm", width=20, height=15)
307
308 #Home Team Handicap v. Mean Consensus P(Home Win)
309 #We'd expect a higher handicap => lower P_cons(Home Win)
310 ensco$AvgHProb.2dp <- with(ensco, round(AvgHProb, 2))
311 ensco$AvgAProb.2dp <- with(ensco, round(AvgAProb, 2))
312
313 ensco$Level <- as.factor(ensco$Level)
314 handicap.v.hprob <- ggplot(ensco, aes(x=AvgHProb.2dp, y=HomeHandicap,
315                                       color=Level)) +
316   geom_count(show.legend = T) +
317   scale_size_area() + theme_light() +
318   labs(x="Consensus Probability of a Home Win (1X2)",
319        y="Home Handicap",
320        title="Plot of Bookmaker Cons. P(Home Win)\nv. Home Handicap",
321        caption="English and Scottish Leagues, 2005-2020")
322
323
324 handicap.v.1x2 <- ggplot(ensco, aes(x=AvgAProb.2dp, y=HomeHandicap,
325                                     color="Away Win")) +
326   geom_count(alpha=.5) +
327   scale_size_area() + theme_light() +
328   geom_count(mapping=aes(x=AvgHProb.2dp, y=HomeHandicap,
329                          color="Home Win"), alpha=.5) +
330   labs(x="Consensus Probability", y="Home Handicap",
331        title="Plot of Bookmaker Cons. of Win (1X2)\nv. Home Handicap",
332        caption="English and Scottish Leagues, 2005-2020") +
333   scale_color_manual(name = "Bet", values = c("Home Win" = "blue",
334                                               "Away Win" = "coral"))
335
336 ggsave(path = "./writeup/images", filename = "ensco_05_hprob_v_hcap.png",
337       plot=handicap.v.hprob, unit="cm", width=15, height=10)
338 ggsave(path = "./writeup/images", filename = "ensco_05a_1x2_v_hcap.png",
339       plot=handicap.v.1x2, unit="cm", width=15, height=10)
340
341
342 #Tile Plots for 1X2 and UO Markets
343 #Our highest-scoring draw was 6-6: As before, we bin 6+ (not 5) goals
    together:
344 N = nrow(ensco); ensco$FTHG.Tile<-with(ensco,rep(0,N))
345 ensco$FTAG.Tile<-with(ensco,rep(0,N))
346
347 for (k in 1:N){
348   if ((ensco$FTHG[k])>=6){ensco$FTHG.Tile[k]<-6}
349   else{ensco$FTHG.Tile[k]<-ensco$FTHG[k]}}
350 for (k in 1:N){
351   if ((ensco$FTAG[k])>=6){ensco$FTAG.Tile[k]<-6}
352   else{ensco$FTAG.Tile[k]<-ensco$FTAG[k]}}
353
354 tile.1x2 <- ggplot(ensco, aes(y=FTAG.Tile, x=FTHG.Tile)) +
355   geom_tile(aes(fill = Correct1X2)) +
356   scale_fill_distiller(palette = "Greens", direction = 1,
357                        name="Correct 1X2\nProbability") +
358   theme_light() +
```

```
  labs(title="Result v. Correct Cons.\nProbability in the 1X2 Market",
       x="Home Goals", y="Away Goals",
       caption="English and Scottish Leagues, 2005-2020") +
  scale_y_discrete(limits=factor(c(1:5, "6+"))) +
  scale_x_discrete(limits=factor(c(1:5, "6+"))) +
  geom_abline(intercept=0, slope=1) +
  coord_cartesian(xlim=c(0,6), ylim=c(0,6))

tile.uo <- ggplot(ensco, aes(y=FTAG.Tile, x=FTHG.Tile)) +
  geom_tile(aes(fill = Over2.5Prob)) +
  scale_fill_distiller(palette = "Paired", direction = 1,
                       name="P(Over 2.5 Goals)") +
  theme_light() +
  labs(title="Result v. Cons.\nProbability of Over 2.5 Goals",
       x="Home Goals", y="Away Goals",
       caption="English and Scottish Leagues, 2005-2020") +
  scale_y_discrete(limits=factor(c(1:5, "6+"))) +
  scale_x_discrete(limits=factor(c(1:5, "6+"))) +
  coord_cartesian(xlim=c(0,6), ylim=c(0,6)) +
  geom_segment(aes(x=2.5,xend=2.5,y=-5,yend=2.5),color="black") +
  geom_segment(aes(x=-5,xend=2.5,y=2.5,yend=2.5),color="black")

ggsave(path = "./writeup/images", filename = "ensco_06_tile_1x2.png",
       plot=tile.1x2, unit="cm", width=15, height=10)
ggsave(path = "./writeup/images", filename = "ensco_07_tile_uo.png",
       plot=tile.uo, unit="cm", width=15, height=10)

tpbinsizes.ensco <- table(ensco$FTAG.Tile, ensco$FTHG.Tile) #Bin sizes

##Under/Over v. Handicap
#Are higher handicap games (more 'obvious') likely to have higher goals?
ensco$Over2.5Prob.2dp <- with(ensco, round(Over2.5Prob, 2))
handicap.v.over2.5 <- ggplot(ensco,
                             aes(x=Over2.5Prob.2dp, y=HomeHandicap)) +
  geom_count() + scale_size_area() + theme_light() +
  labs(x="Consensus Probability of a Over 2.5 Goals", y="Home Handicap",
       title="Plot of Bookmaker Cons. P(Over 2.5 Goals)\nv. HHandicap",
       caption="English and Scottish Leagues, 2005-2020")

ensco$GoalDiffTile <- with(ensco, FTAG.Tile-FTHG.Tile)
expected.v.act.difference <- ggplot(ensco, aes(y=GoalDiffTile,
                                               x=HomeHandicap)) +
  geom_count() + scale_size_area() + theme_light() +
  labs(x="Home Handicap (Expected GD)",
       y="Away Goals minus Home Goals (Actual GD)",
       title="Plot of Expected v. Actual Goal Difference",
       caption="English and Scottish Leagues, 2005-2020")

ggsave(path = "./writeup/images",
       filename = "ensco_08_handicap_v_over.png",
       plot=handicap.v.over2.5, unit="cm", width=15, height=10)
ggsave(path = "./writeup/images",
       filename = "ensco_09_exp_v_act_goaldiff.png",
       plot=expected.v.act.difference, unit="cm", width=15, height=10)
```

```
413
414 ### CORRELATION TESTS (Kendall's Tau, Spearman) ----
415 cor.test(ensco$FTHG.Tile, ensco$FTAG.Tile, method = "kendall")
416 cor.test(ensco$FTHG.Tile, ensco$FTAG.Tile, method = 'spearman')
417 #p << 0.005: Strong evidence of an association. => ???
418 library(DescTools)
419 GoodmanKruskalGamma(ensco$FTHG.Tile, ensco$FTAG.Tile, conf.level = 0.95)
420
421
422 ### CORRELATION ANALYSIS: MODEL CREATION (Overall) ----
423
424 #Cutting and defining the levels:
425 ensco$AvgHProb.cut <- cut(ensco$AvgHProb, 50, include.lowest=T)
426 levels(ensco$AvgHProb.cut) <- tapply(ensco$AvgHProb,
427                                      ensco$AvgHProb.cut, mean)
428 ensco$AvgDProb.cut <- cut(ensco$AvgDProb, 50, include.lowest=T)
429 levels(ensco$AvgDProb.cut) <- tapply(ensco$AvgDProb,
430                                      ensco$AvgDProb.cut, mean)
431 ensco$AvgAProb.cut <- cut(ensco$AvgAProb, 50, include.lowest=T)
432 levels(ensco$AvgAProb.cut) <- tapply(ensco$AvgAProb,
433                                      ensco$AvgAProb.cut, mean)
434
435 ensco$Over2.5Prob.cut <- cut(ensco$Over2.5Prob, 50, include.lowest=T)
436 levels(ensco$Over2.5Prob.cut) <- tapply(ensco$Over2.5Prob,
437                                         ensco$Over2.5Prob.cut, mean)
438
439 ensco$AH.HProb.cut <- cut(ensco$AH.HProb, 50, include.lowest=T)
440 levels(ensco$AH.HProb.cut) <- tapply(ensco$AH.HProb,
441                                      ensco$AH.HProb.cut, mean)
442 ensco$AH.AProb.cut <- cut(ensco$AH.AProb, 50, include.lowest=T)
443 levels(ensco$AH.AProb.cut) <- tapply(ensco$AH.AProb,
444                                      ensco$AH.AProb.cut, mean)
445
446 #Observed Probability for each cut:
447 obsprob.1x2.H <- prop.table(table(ensco$FTR, ensco$AvgHProb.cut), 2)[3,]
448 booprob.1x2.H <- as.numeric(names(obsprob.1x2.H))
449 obsprob.1x2.D <- prop.table(table(ensco$FTR, ensco$AvgDProb.cut), 2)[2,]
450 booprob.1x2.D <- as.numeric(names(obsprob.1x2.D))
451 obsprob.1x2.A <- prop.table(table(ensco$FTR, ensco$AvgAProb.cut), 2)[1,]
452 booprob.1x2.A <- as.numeric(names(obsprob.1x2.A))
453
454 booprob.1x2 <- c(booprob.1x2.H,booprob.1x2.D,booprob.1x2.A)
455 obsprob.1x2 <- c(obsprob.1x2.H,obsprob.1x2.D,obsprob.1x2.A)
456
457 obsprob.uo <- prop.table(table(ensco$uo.res,
458                                ensco$Over2.5Prob.cut), 2)[1,]
459 booprob.uo <- as.numeric(names(obsprob.uo))
460
461 #For AH bets, we will only take full wins:
462 obsprob.ah.H <- prop.table(table(ensco$ah.res,
463                                  ensco$AH.HProb.cut), 2)[4,] #4 = Home
464 booprob.ah.H <- as.numeric(names(obsprob.ah.H))
465 obsprob.ah.A <- prop.table(table(ensco$ah.res,
466                                  ensco$AH.AProb.cut), 2)[1,] #1 = Away
```

```
booprob.ah.A <- as.numeric(names(obsprob.ah.A))

booprob.ah <- c(booprob.ah.H, booprob.ah.A)
obsprob.ah <- c(obsprob.ah.H, obsprob.ah.A)

#Final models
model.1x2.h <- lm(obsprob.1x2.H~booprob.1x2.H)
model.1x2.d <- lm(obsprob.1x2.D~booprob.1x2.D)
model.1x2.a <- lm(obsprob.1x2.A~booprob.1x2.A)
model.1x2.o <- lm(obsprob.1x2~booprob.1x2)
model.uo <- lm(obsprob.uo~booprob.uo)
model.ah <- lm(obsprob.ah~booprob.ah)

#R Squared and RMSE values:-
rsrm.val.1x2 <- matrix(c(summary(model.1x2.h)$r.squared,
                         summary(model.1x2.d)$r.squared,
                         summary(model.1x2.a)$r.squared,
                         sqrt(mean(model.1x2.h$residuals^2)),
                         sqrt(mean(model.1x2.d$residuals^2)),
                         sqrt(mean(model.1x2.a$residuals^2)) ),
                       ncol = 3, nrow = 2, byrow=T,
                       dimnames = list(c("RSq", "RMSE"),
                                       c("1X2: H", "D", "A")))
rsrm.val.uoah <- matrix(c(summary(model.uo)$r.squared,
                          summary(model.ah)$r.squared,
                         sqrt(mean(model.uo$residuals^2)),
                         sqrt(mean(model.ah$residuals^2)) ),
                        ncol = 2, nrow = 2, byrow=T,
                        dimnames = list(c("RSq", "RMSE"),
                                        c("Under/Over", "AH")))

#These show that the odds for the U/O market, 1x2 Draws are NOT accurate

#    Model Plots ----
convobs.1x2 <- ggplot(data=NULL, aes()) + geom_smooth() +
  geom_jitter(aes(x=booprob.1x2.H,
                  y=obsprob.1x2.H, color="1X2 Home"), size=0.75) +
  geom_smooth(aes(x=booprob.1x2.H,
                  y=obsprob.1x2.H, color="1X2 Home"), method=lm) +
  geom_jitter(aes(x=booprob.1x2.D,
                  y=obsprob.1x2.D, color="1X2 Draw"), size=0.75) +
  geom_smooth(aes(x=booprob.1x2.D,
                  y=obsprob.1x2.D, color="1X2 Draw"), method=lm) +
  geom_jitter(aes(x=booprob.1x2.A,
                  y=obsprob.1x2.A, color="1X2 Away"), size=0.75) +
  geom_smooth(aes(x=booprob.1x2.A,
                  y=obsprob.1x2.A, color="1X2 Away"), method=lm) +
  geom_abline(intercept=0, slope=1, linetype="dashed") + theme_light() +
  labs(title="Cons v. Obs Probabilities - 1x2 Market",
       caption="Eng/Sco 05-20") +
  scale_color_manual(name="Bet Type",
                     values=c("1X2 Home" = "blue", "1X2 Draw" = "green4",
                              "1X2 Away" = "coral")) +
  coord_cartesian(xlim=c(0,1), ylim=c(0,1))
```

```
521
522 convobs.uo <- ggplot(data=NULL, aes()) + geom_smooth() +
523   geom_jitter(aes(x=booprob.uo,
524                   y=obsprob.uo, color="Under/Over"), size=0.75) +
525   geom_smooth(aes(x=booprob.uo,
526                   y=obsprob.uo, color="Under/Over"), method=lm) +
527   geom_abline(intercept=0, slope=1, linetype="dashed") + theme_light() +
528   labs(title="Cons v. Obs Probabilities - Under/Over Market",
529        caption="Eng/Sco 05-20") +
530   scale_color_manual(name="Bet Type", values=c("Under/Over" = "red")) +
531   coord_cartesian(xlim=c(0,1), ylim=c(0,1))
532
533 convobs.ah <- ggplot(data=NULL, aes()) + geom_smooth() +
534   geom_jitter(aes(x=booprob.ah.H,
535                   y=obsprob.ah.H, color="AH Home"), size=0.75) +
536   geom_smooth(aes(x=booprob.ah.H,
537                   y=obsprob.ah.H, color="AH Home"), method=lm) +
538   geom_jitter(aes(x=booprob.ah.A,
539                   y=obsprob.ah.A, color="AH Away"), size=0.75) +
540   geom_smooth(aes(x=booprob.ah.A,
541                   y=obsprob.ah.A, color="AH Away"), method=lm) +
542   geom_abline(intercept=0, slope=1, linetype="dashed") + theme_light() +
543   labs(title="Cons v. Obs Probabilities - Asian Handicap Market",
544        caption="Eng/Sco 05-20") +
545   scale_color_manual(name="Bet Type", values=c("AH Home" = "blue",
546                                                "AH Away" = "coral")) +
547   coord_cartesian(xlim=c(0,1), ylim=c(0,1))
548
549 ggsave(path = "./writeup/images", filename = "ensco_10_convobs_1x2.png",
550       plot=convobs.1x2, unit="cm", width=15, height=6)
551 ggsave(path = "./writeup/images", filename = "ensco_11_convobs_uo.png",
552       plot=convobs.uo, unit="cm", width=15, height=6)
553 ggsave(path = "./writeup/images", filename = "ensco_12_convobs_ah.png",
554       plot=convobs.ah, unit="cm", width=15, height=6)
555
556 ### CORRELATION ANALYSIS: PER LEVEL ----
557 #To view sample size:-
558 for (j in levels){
559   DataTemp <- ensco[ensco$Level==j,]
560   print(paste0("For level ",j,", n = ",nrow(DataTemp)))
561 }
562 rsqu.level <- NULL; rmse.level <- NULL; rsqu.level.1x2 <- NULL;
563 rmse.level.1x2 <- NULL; rsqu.level.uo <- NULL; rmse.level.uo <- NULL;
564 rsqu.level.ah <- NULL; rmse.level.ah <- NULL; p1.level <- NULL
565 p2.level <- NULL
566 ensco$LogCorrect1x2 <- with(ensco, log(ensco$Correct1X2))
567
568 #n.b. We don't do a model for each -- just an overall 1x2, AH and U/O.
569 #     P1 and P2 is based purely on the 1X2 market.
570
571 for (j in levels){
572   dataTemp <- ensco[ensco$Level==j,]
573   dataTemp$AvgHProb.cut <-
574     cut(dataTemp$AvgHProb, 35, include.lowest = T)
```

```
575   levels(dataTemp$AvgHProb.cut) <-
576     tapply(dataTemp$AvgHProb, dataTemp$AvgHProb.cut, mean)
577   dataTemp$AvgDProb.cut <-
578     cut(dataTemp$AvgDProb, 15, include.lowest = T)
579   levels(dataTemp$AvgDProb.cut) <-
580     tapply(dataTemp$AvgDProb, dataTemp$AvgDProb.cut, mean)
581   dataTemp$AvgAProb.cut <-
582     cut(dataTemp$AvgAProb, 35, include.lowest = T)
583   levels(dataTemp$AvgAProb.cut) <-
584     tapply(dataTemp$AvgAProb, dataTemp$AvgAProb.cut, mean)
585
586   dataTemp$Over2.5Prob.cut <-
587     cut(dataTemp$Over2.5Prob, 35, include.lowest = T)
588   levels(dataTemp$Over2.5Prob.cut) <-
589     tapply(dataTemp$Over2.5Prob, dataTemp$Over2.5Prob.cut, mean)
590
591   dataTemp$AH.HProb.cut <-
592     cut(dataTemp$AH.HProb, 35, include.lowest = T)
593   levels(dataTemp$AH.HProb.cut) <-
594     tapply(dataTemp$AH.HProb, dataTemp$AH.HProb.cut, mean)
595   dataTemp$AH.AProb.cut <-
596     cut(dataTemp$AH.AProb, 35, include.lowest = T)
597   levels(dataTemp$AH.AProb.cut) <-
598     tapply(dataTemp$AH.AProb, dataTemp$AH.AProb.cut, mean)
599
600   obs.1x2.h <- prop.table(table(dataTemp$FTR,
601                                 dataTemp$AvgHProb.cut), 2)[3,]
602   obs.1x2.d <- prop.table(table(dataTemp$FTR,
603                                 dataTemp$AvgDProb.cut), 2)[2,]
604   obs.1x2.a <- prop.table(table(dataTemp$FTR,
605                                 dataTemp$AvgAProb.cut), 2)[1,]
606   obs.1x2 <- c(obs.1x2.h, obs.1x2.d, obs.1x2.a)
607
608   boo.1x2.h <- as.numeric(names(obs.1x2.h))
609   boo.1x2.d <- as.numeric(names(obs.1x2.d))
610   boo.1x2.a <- as.numeric(names(obs.1x2.a))
611   boo.1x2 <- c(boo.1x2.h, boo.1x2.d, boo.1x2.a)
612
613   obs.uo <- prop.table(table(dataTemp$uo.res,
614                              dataTemp$Over2.5Prob.cut), 2)[1,]
615   boo.uo <- as.numeric(names(obs.uo))
616
617   obs.ah.h <- prop.table(table(dataTemp$ah.res,
618                                dataTemp$AH.HProb.cut), 2)[4,]
619   obs.ah.a <- prop.table(table(dataTemp$ah.res,
620                                dataTemp$AH.AProb.cut), 2)[1,]
621   obs.ah <- c(obs.ah.h, obs.ah.a)
622
623   boo.ah.h <- as.numeric(names(obs.ah.h))
624   boo.ah.a <- as.numeric(names(obs.ah.a))
625   boo.ah <- c(boo.ah.h, boo.ah.a)
626
627   modelTemp.1x2 <- lm(obs.1x2 ~ boo.1x2)
628   modelTemp.uo <- lm(obs.uo ~ boo.uo)
```

```
629  modelTemp.ah <- lm(obs.ah ~ boo.ah)
630  par(mfrow=c(3,1))
631  plot(modelTemp.1x2, 5, main = paste0("1X2 Market; Level ",j))
632  plot(modelTemp.uo, 5, main = paste0("UO Market; Level ",j))
633  plot(modelTemp.ah, 5, main = paste0("AH Market; Level ",j))
634  par(mfrow=c(1,1))
635  plot.1x2 <- ggplot(NULL, aes(x=boo.1x2.h, y=obs.1x2.h, color="Home")) +
636    geom_smooth(method="lm", alpha=0.3) +
637    geom_smooth(aes(x=boo.1x2.a, y=obs.1x2.a, color="Away"),
638                method="lm", alpha=0.3) +
639    geom_smooth(aes(x=boo.1x2.d, y=obs.1x2.d, color="Draw"),
640                method="lm", alpha=0.3) +
641    geom_jitter(aes(color="Home"), shape=1) +
642    geom_jitter(aes(x=boo.1x2.a, y=obs.1x2.a, color="Away"), shape=1) +
643    geom_jitter(aes(x=boo.1x2.d, y=obs.1x2.d, color="Draw"), shape=1) +
644    geom_abline(slope=1, intercept=0, color="black", linetype="dashed") +
645    coord_cartesian(xlim=c(0,1), ylim=c(0,1)) +
646    labs(x = paste0("1X2 Bookmaker Consensus Probabilities: Level ",j),
647         y = NULL) +
648    scale_color_manual(name="Bet Type",
649                       values = c("Home" = "blue", "Away" = "coral",
650                                  "Draw" = "green4")) + theme_light()
651
652  plot.uo <- ggplot(NULL, aes(x=boo.uo, y=obs.uo,
653                              color="Over 2.5 Goals")) +
654    geom_smooth(method="lm", alpha=0.3) + geom_jitter(shape=2) +
655    geom_abline(slope=1, intercept=0, color="black", linetype="dashed") +
656    coord_cartesian(xlim=c(0,1), ylim=c(0,1)) +
657    labs(x=paste0("UO Bookmaker Consensus Probabilities: Level ",j),
658         y=NULL) +
659    scale_color_manual(name="Bet Type",
660                       values=c("Over 2.5 Goals" = "red")) +
661    theme_light()
662
663  plot.ah <- ggplot(NULL, aes(x=boo.ah.h, y=obs.ah.h, color="AH Home")) +
664    geom_smooth(method = "lm", alpha=0.3) +
665    geom_smooth(aes(x=boo.ah.a, y=obs.ah.a, color="AH Away"),
666                method="lm", alpha=0.3) + geom_jitter(shape=5) +
667    geom_jitter(aes(x=boo.ah.a, y=obs.ah.a, color="AH Away"),
668                method="lm", alpha=0.3) +
669    geom_abline(slope=1, intercept=0, color="black", linetype="dashed") +
670    coord_cartesian(xlim=c(0,1), ylim=c(0,1)) +
671    labs(x= paste0("AH Bookmaker Consensus Probabilities: Level ",j),
672         y=NULL) +
673    scale_color_manual(name="Bet Type",
674                       values=c("AH Home"="blue", "AH Away"="coral"))+
675    theme_light()
676
677
678  plotTemp <- grid.arrange(plot.1x2, plot.uo, plot.ah, nrow=3, ncol=1,
679                           left="Observed Probability")
680  ggsave(path = "./writeup/images",
681         filename = paste0("ensco_13_level",j,".png"), plot=plotTemp,
682         unit="cm", width=15, height=10)
```

```
683
684
685   rsqu.level.1x2 <- c(rsqu.level.1x2,
686                       round(summary(modelTemp.1x2)$r.squared, 5))
687   rmse.level.1x2 <- c(rmse.level.1x2,
688                       round(sqrt(mean(modelTemp.1x2$residuals^2)), 5))
689   rsqu.level.uo <- c(rsqu.level.uo,
690                      round(summary(modelTemp.uo)$r.squared, 5))
691   rmse.level.uo <- c(rmse.level.uo,
692                      round(sqrt(mean(modelTemp.uo$residuals^2)), 5))
693   rsqu.level.ah <- c(rsqu.level.ah,
694                      round(summary(modelTemp.ah)$r.squared, 5))
695   rmse.level.ah <- c(rmse.level.ah,
696                      round(sqrt(mean(modelTemp.ah$residuals^2)), 5))
697
698   p1.temp <- exp(1/(nrow(dataTemp)) * sum(dataTemp$LogCorrect1x2))
699   p2.temp <- 1/(nrow(dataTemp))*sum((1 - dataTemp$Correct1X2)**2 +
700                                     (dataTemp$IncorrectA1X2)**2 +
701                                     (dataTemp$IncorrectB1X2)**2 )
702   p1.level <- c(p1.level, round(p1.temp, 5))
703   p2.level <- c(p2.level, round(p2.temp ,5))
704 }
705 rmse.level <- matrix( c(rmse.level.1x2, rmse.level.uo, rmse.level.ah),
706                       ncol=3, byrow=T,
707                       dimnames = list(c("1x2", "UO", "AH"), levels))
708 rsqu.level <- matrix( c(rsqu.level.1x2, rsqu.level.uo, rsqu.level.ah),
709                       ncol=3, byrow=T,
710                       dimnames = list(c("1x2", "UO", "AH"), levels))
711 p.values.level <- matrix(c(p1.level, p2.level), ncol=3, byrow=T,
712                          dimnames = list(c("P1", "P2"), levels))
713
714 ### OVERROUND PLOTS ----
715 or.ot.l <- ggplot(ensco, aes(x=AvgHProb, y=OneXTwoOverround, color =
        Level)) +
716   geom_jitter(alpha = 0.5) + theme_light() +
717   guides(col = guide_legend(ncol = 3)) +
718   labs(x = "Consensus P(Home Win)", y = "Sum of Probabilities (1X2)",
          title =
719   "Consensus P(Home Win) v. Bookmaker Commission, by Level",
720   caption = "English/Scottish Leagues, 2005-20")
721
722 or.ot.s <- ggplot(ensco, aes(x=AvgHProb, y=OneXTwoOverround, color =
        Season)) +
723   geom_jitter(alpha = 0.5) + theme_light() +
724   guides(col = guide_legend(ncol = 3)) +
725   labs(x = "Consensus P(Home Win)", y = "Sum of Probabilities (1X2)",
          title =
726     "Consensus P(Home Win) v. Bookmaker Commission, by Season",
727    caption = "English/Scottish Leagues, 2005-20")
728
729 or.uo.l <-
        ggplot(ensco,aes(x=Over2.5Prob,y=UnderOverOverround,color=Level)) +
730   geom_jitter(alpha = 0.5) + theme_light() +
731   guides(col = guide_legend(ncol = 3)) +
```

```
732   labs(x = "Consensus P(Over 2.5 Goals)", y="Sum of Probabilities
          (UO)",title =
733         "Consensus P(Over 2.5 Goals) v. Bookmaker Commission, by Level",
734       caption = "English/Scottish Leagues, 2005-20")
735
736
737 or.uo.s <-
        ggplot(ensco,aes(x=Over2.5Prob,y=UnderOverOverround,color=Season)) +
738   geom_jitter(alpha = 0.5) + theme_light() +
739   guides(col = guide_legend(ncol = 3)) +
740   labs(x = "Consensus P(Over 2.5 Goals)", y="Sum of Probabilities
          (UO)",title =
741         "Consensus P(Over 2.5 Goals) v. Bookmaker Commission, by
              Season",
742       caption = "English/Scottish Leagues, 2005-20")
743
744 or.ah.l <- ggplot(ensco, aes(x=AH.HProb, y=AHOverround, color = Level)) +
745   geom_jitter(alpha = 0.5) + theme_light() +
746   guides(col = guide_legend(ncol = 3)) +
747   labs(x = "Consensus P(AH Home Win)", y = "Sum of Probabilities (AH)",
          title =
748         "Consensus P(Home Win, AH) v. Bookmaker Commission, by Level",
749       caption = "English/Scottish Leagues, 2005-20")
750
751 or.ah.s <- ggplot(ensco, aes(x=AH.HProb, y=AHOverround, color = Season))
        +
752   geom_jitter(alpha = 0.5) + theme_light() +
753   guides(col = guide_legend(ncol = 3)) +
754   labs(x = "Consensus P(AH Home Win)", y = "Sum of Probabilities (AH)",
          title =
755         "Consensus P(Home Win, AH) v. Bookmaker Commission, by Season",
756       caption = "English/Scottish Leagues, 2005-20")
757
758 ggsave(path = "./writeup/images", filename =
        "ensco_14a_overround_ot_l.png",
759       plot=or.ot.l, unit="cm", width=20, height=10)
760 ggsave(path = "./writeup/images", filename =
        "ensco_14b_overround_ot_s.png",
761       plot=or.ot.s, unit="cm", width=20, height=10)
762 ggsave(path = "./writeup/images", filename =
        "ensco_14c_overround_uo_l.png",
763       plot=or.uo.l, unit="cm", width=20, height=10)
764 ggsave(path = "./writeup/images", filename =
        "ensco_14d_overround_uo_s.png",
765       plot=or.uo.s, unit="cm", width=20, height=10)
766 ggsave(path = "./writeup/images", filename =
        "ensco_14e_overround_ah_l.png",
767       plot=or.ah.l, unit="cm", width=20, height=10)
768 ggsave(path = "./writeup/images", filename =
        "ensco_14f_overround_ah_s.png",
769       plot=or.ah.s, unit="cm", width=20, height=10)
770
771 #Overround Calcs
772 #Overall, across all leagues/season:
```

```
mean(ensco$OneXTwoOverround)
mean(ensco$UnderOverOverround)
mean(ensco$AHOverround)
#
#By Level
for (i in levels){
  print(paste("Level",i,"-------------------------------------------------"))
  print(paste("1X2 Overround:
      ",mean(ensco[ensco$Level==i,]$OneXTwoOverround)))
  print(paste("UO Overround:
      ",mean(ensco[ensco$Level==i,]$UnderOverOverround)))
  print(paste("AH Overround: ",mean(ensco[ensco$Level==i,]$AHOverround)))
}
#Select seasons (05/06, 12/13 and 19/20) (equally spaced)
for (j in c("0506", "1213", "1920")){
  print(paste("Season",j,"-------------------------------------------------"))
  print(paste("1X2 Overround:
      ",mean(ensco[ensco$Season==j,]$OneXTwoOverround)))
  print(paste("UO Overround:
      ",mean(ensco[ensco$Season==j,]$UnderOverOverround)))
  print(paste("AH Overround:
      ",mean(ensco[ensco$Season==j,]$AHOverround)))
}

#### IMPORTANT OUTPUTS ----
#Plots (if ggsave isn't used, the commands to view the plots):
  #Density plots of probability distributions:
#       dens1x2ha; dens1x2d; densuo; densah
  #Other visual analysis plots:
#       hprob.v.handicap; tile.1x2; tile.uo; handicap.v.over2.5;
#       expected.v.actual.difference
  #Consensus v. Observed:
#       convobs.1x2; convobs.uo; convobs.ah
  #Winnings:
#       cumulwinningsplot; cumulwinningsplot100

#Matrices and Tables
  basic.calcs; obsprob.1x2tab; obsprob.uotab; obsprob.ahtab
  rsqu.level; rmse.level; p.values.level
  tpbinsizes.ensco
  for (j in levels){
    DataTemp <- ensco[ensco$Level==j,]
    print(paste0("For level ",j,", n = ",nrow(DataTemp)))
  }
```

Appendix D

Chapter 4 Code

```
#### (3) USING OUR RESULTS FOR A BETTING METHOD
library(ggplot2); library(scales); library(e1071)

## CLEANING AND OBTAINING DATA ----
# First, we need to trim the datasets to just the columns we need and
# obtain the Asian Handicap odds for elite leagues from F-D.co.uk

matchesTemp <- NULL; matches <- NULL
seasons <- c("0506", "0607", "0708", "0809", "0910", "1011", "1112",
             "1213", "1314", "1415", "1516", "1617", "1718", "1819",
             "1920")
divisions <- c("E0", "E1", "E2", "E3", "EC", "SC0", "SC1", "SC2", "SC3",
    "D1",
               "SP1", "F1", "I1", "P1")

for (i in seasons){
  for (j in divisions){
    matchesTemp <- read.csv(paste0(
      "https://www.football-data.co.uk/mmz4281/",i,"/",j,".csv"),
      fileEncoding="latin1")
    #The above line will download and read the .csv file in one go.
    matchesTemp$Season <- with(matchesTemp,i)
    matchesTemp$Div <- with(matchesTemp, j)
    if (i=="1920"){
      matchesTemp$BbAvH <- matchesTemp$AvgH
      matchesTemp$BbAvA <- matchesTemp$AvgA
      matchesTemp$BbAvD <- matchesTemp$AvgD
      matchesTemp$BbAvAHH <- matchesTemp$AvgAHH
      matchesTemp$BbAvAHA <- matchesTemp$AvgAHA
      matchesTemp$BbAHh <- matchesTemp$AHh}
    else{}
    matchesTemp$HomeHandicap <- matchesTemp$BbAHh
    matchesTemp <- matchesTemp[,c("Div", "Date", "HomeTeam", "AwayTeam",
                            "FTHG", "FTAG", "FTR", "BbAvH", "BbAvD",
                            "BbAvA", "HomeHandicap", "BbAvAHH", "BbAvAHA",
                            "Season")]
    matches <- rbind(matches, matchesTemp)
```

```
  }
}
matches <- na.omit(matches)

# Finding underlying probabilities of each event:-
matches$OT.HProb.PN <- with(matches, 1/(BbAvH))
matches$OT.DProb.PN <- with(matches, 1/(BbAvD))
matches$OT.AProb.PN <- with(matches, 1/(BbAvA))
matches$AH.HProb.PN <- with(matches, 1/(BbAvAHH))
matches$AH.AProb.PN <- with(matches, 1/(BbAvAHA))

matches$OT.HProb <-
  with(matches, round(OT.HProb.PN/(OT.HProb.PN+OT.DProb.PN+OT.AProb.PN),
      4))
matches$OT.AProb <-
  with(matches, round(OT.AProb.PN/(OT.HProb.PN+OT.DProb.PN+OT.AProb.PN),
      4))
matches$AH.HProb <-
    with(matches,round(AH.HProb.PN/(AH.HProb.PN+AH.AProb.PN),4))
matches$AH.AProb <-
    with(matches,round(AH.AProb.PN/(AH.HProb.PN+AH.AProb.PN),4))

N = nrow(matches)

matches$FTHG.ah <- with(matches, rep(0,N))
for (m in 1:N){matches$FTHG.ah[m] <- matches$FTHG[m] +
    matches$HomeHandicap[m]}
matches$ah.gap <- with(matches, FTHG.ah - FTAG); matches$ah.res <- NULL
for (n in 1:N){
  if (matches$ah.gap[n]<(-0.25)){matches$ah.res[n]<-"aw"}
  else if (matches$ah.gap[n]==(-0.25)){matches$ah.res[n]<-"hfaw"}
  else if (matches$ah.gap[n]==0){matches$ah.res[n]<-"vo"}
  else if (matches$ah.gap[n]==0.25){matches$ah.res[n]<-"hfhm"}
  else if (matches$ah.gap[n]>0.25){matches$ah.res[n]<-"hm"}
  else{}
}


## PLACING BETS ----
# To assess our method, we assume we use the first year (05/06) to obtain
# data (that is, the means and std devs) and place bets on any game
    after.
matches$OTHomeBet <- with(matches, 0); matches$OTAwayBet <-
    with(matches, 0)
matches$AHHomeBet <- with(matches, 0); matches$AHAwayBet <-
    with(matches, 0)

#Initial Bounds :-
matches0506 <- matches[matches$Season == '0506',]
mu.oth <- mean(matches0506$OT.HProb); sd.oth <- sd(matches0506$OT.HProb)
mu.ota <- mean(matches0506$OT.AProb); sd.ota <- sd(matches0506$OT.AProb)
mu.ahh <- mean(matches0506$AH.HProb); sd.ahh <- sd(matches0506$AH.HProb)
mu.aha <- mean(matches0506$AH.AProb); sd.aha <- sd(matches0506$AH.AProb)
n <- nrow(matches[matches$Season == "0506",]);N <- nrow(matches)
```

```
#Placing Bets:-
for (i in n:N){
  #Update the mean and std dev's with our new information
  mu.oth <- mean(matches$OT.HProb[1:i]); sd.oth <-
      sd(matches$OT.HProb[1:i])
  mu.ota <- mean(matches$OT.AProb[1:i]); sd.ota <-
      sd(matches$OT.AProb[1:i])
  mu.ahh <- mean(matches$AH.HProb[1:i]); sd.ahh <-
      sd(matches$AH.HProb[1:i])
  mu.aha <- mean(matches$AH.AProb[1:i]); sd.aha <-
      sd(matches$AH.AProb[1:i])
  #Do we bet on Home Win (1X2)?
  if (matches$OT.HProb[i] > mu.oth + 0.5*sd.oth){
    if (matches$OT.HProb[i] <= mu.oth + sd.oth){matches$OTHomeBet[i] <-
        1}
    else if (matches$OT.HProb[i] > mu.oth + sd.oth &
             matches$OT.HProb[i] <= mu.oth + 1.5*sd.oth){
      matches$OTHomeBet[i] <- 2}
    else {matches$OTHomeBet[i] <- 3}}
  else {matches$OTHomeBet[i] <- 0}
  #Do we bet on Away Win (1X2)?
  if (matches$OT.AProb[i] > mu.ota + 0.5*sd.ota){
    if (matches$OT.AProb[i] <= mu.ota + sd.ota){matches$OTAwayBet[i] <-
        1}
    else if (matches$OT.AProb[i] > mu.ota + sd.ota &
             matches$OT.AProb[i] <= mu.ota + 1.5*sd.ota){
      matches$OTAwayBet[i] <- 2}
    else {matches$OTAwayBet[i] <- 3}}
  else {matches$OTAwayBet[i] <- 0}
  #Do we bet on Home Win (AH)?
  if (matches$AH.HProb[i] > mu.ahh + sd.ahh){
    if (matches$AH.HProb[i] <= mu.ahh + 1.5*sd.ahh){matches$AHHomeBet[i]
        <- 1}
    else if (matches$AH.HProb[i] > mu.ahh + 1.5*sd.ahh &
             matches$AH.HProb[i] <= mu.ahh + 2*sd.ahh){
      matches$AHHomeBet[i] <- 2}
    else {matches$AHHomeBet[i] <- 3}}
  else {matches$AHHomeBet[i] <- 0}
  #Do we bet on Away Win (AH)?
  if (matches$AH.AProb[i] > mu.aha + sd.aha){
    if (matches$AH.AProb[i] <= mu.aha + 1.5*sd.aha){matches$AHAwayBet[i]
        <- 1}
    else if (matches$AH.AProb[i] > mu.aha + 1.5*sd.aha &
             matches$AH.AProb[i] <= mu.aha + 2*sd.aha){
      matches$AHAwayBet[i] <- 2}
    else {matches$AHAwayBet[i] <- 3}}
  else {matches$AHAwayBet[i] <- 0}
}

## FINDING RESULTS ----
matches$OTHReturns <- with(matches, 0); matches$OTAReturns <-
    with(matches, 0)
matches$AHHReturns <- with(matches, 0); matches$AHAReturns <-
    with(matches, 0)
```

```
127
128 for (i in n:N){
129   if (matches$FTR[i]=="H"){
130     matches$OTHReturns[i] <- (matches$BbAvH[i]-1) * matches$OTHomeBet[i]
131     matches$OTAReturns[i] <- -matches$OTAwayBet[i]}
132   else if (matches$FTR[i]=="A"){
133     matches$OTAReturns[i] <- (matches$BbAvA[i]-1) * matches$OTAwayBet[i]
134     matches$OTHReturns[i] <- -matches$OTHomeBet[i]}
135   else {
136     matches$OTHReturns[i] <- -matches$OTHomeBet[i]
137     matches$OTAReturns[i] <- -matches$OTAwayBet[i]}
138 }
139
140 for (i in n:N){
141   if (matches$ah.res[i]=="aw"){
142     matches$AHAReturns[i] <- (matches$BbAvAHA[i]-1) *
          matches$AHAwayBet[i]
143     matches$AHHReturns[i] <- -matches$AHHomeBet[i]
144   }
145   else if (matches$ah.res[i]=="hfaw"){
146     matches$AHAReturns[i] <-
147       (matches$BbAvAHA[i]-1)*0.5*matches$AHAwayBet[i]
            -0.5*matches$AHAwayBet[i]
148     matches$AHHReturns[i] <- -matches$AHHomeBet[i]
149   }
150   else if (matches$ah.res[i]=="hm"){
151     matches$AHHReturns[i] <- (matches$BbAvAHH[i]-1) *
          matches$AHHomeBet[i]
152     matches$AHAReturns[i] <- -matches$AHAwayBet[i]
153   }
154   else if (matches$ah.res[i]=="hfhm"){
155     matches$AHHReturns[i] <-
156       (matches$BbAvAHH[i]-1)*0.5*matches$AHHomeBet[i]
            -0.5*matches$AHHomeBet[i]
157     matches$AHAReturns[i] <- -matches$AHAwayBet[i]
158   }
159   else{
160     matches$AHHReturns[i] <- -matches$AHHomeBet[i]
161     matches$AHAReturns[i] <- -matches$AHAwayBet[i]
162   }
163 }
164
165 #Cumulative Returns (for our plot):-
166 matches$C.OTHReturns <- with(matches,0); matches$C.OTAReturns <-
      with(matches,0)
167 matches$C.AHHReturns <- with(matches,0); matches$C.AHAReturns <-
      with(matches,0)
168 matches$C.Returns <- with(matches, 0)
169
170 matches$C.OTHReturns[1] <- matches$OTHReturns[1]
171 matches$C.OTAReturns[1] <- matches$OTAReturns[1]
172 matches$C.AHHReturns[1] <- matches$AHHReturns[1]
173 matches$C.AHAReturns[1] <- matches$AHAReturns[1]
174 matches$C.Returns[1] <- matches$OTHReturns[1]+matches$OTAReturns[1]+
```

```
  matches$AHHReturns[1]+matches$AHAReturns[1]

for (i in 2:N){
  matches$C.Returns[i] <-
    matches$C.Returns[i-1]+matches$OTHReturns[i]+matches$OTAReturns[i]+
    matches$AHHReturns[i]+matches$AHAReturns[i]}
for (i in 2:N){
  matches$C.OTHReturns[i] <- matches$C.OTHReturns[i-1] +
      matches$OTHReturns[i]}
for (i in 2:N){
  matches$C.OTAReturns[i] <- matches$C.OTAReturns[i-1] +
      matches$OTAReturns[i]}
for (i in 2:N){
  matches$C.AHHReturns[i] <- matches$C.AHHReturns[i-1] +
      matches$AHHReturns[i]}
for (i in 2:N){
  matches$C.AHAReturns[i] <- matches$C.AHAReturns[i-1] +
      matches$AHAReturns[i]}

matches$BetIndex <- with(matches,0)
for (i in n:N){matches$BetIndex[i] <- (i-(n-1))}

## PLOTS ----
cr1 <- ggplot(matches,aes(x=BetIndex,y=C.OTHReturns,color="1X2 H
    Returns")) +
  geom_line() +
  geom_line(aes(x=BetIndex, y=C.OTAReturns, color="1X2 A Returns")) +
  geom_line(aes(x=BetIndex, y=C.AHHReturns, color="AH H Returns")) +
  geom_line(aes(x=BetIndex, y=C.AHAReturns, color="AH A Returns")) +
  #geom_line(aes(x=BetIndex, y=C.Returns, color = "Total
      Returns"),linetype = "twodash") +
  theme_light() +
  labs(title="Cum. Winnings using our Betting Model",
       caption="All matches 06-20", x="Match Index", y="Returns (unit)")
           +
  scale_color_manual(name="Bet Type",
                     values=c("1X2 H Returns" = "blue",
                              "1X2 A Returns" = "coral",
                              "AH H Returns" = "green",
                              "AH A Returns" = "purple",
                              "Total Returns" = "red")) +
  geom_hline(yintercept=0, color="black", linetype="dotted", alpha=.5)

cr2 <- ggplot(matches, aes(x=BetIndex, y=C.OTHReturns, color="1X2 H
    Returns")) +
  geom_line() +
  geom_line(aes(x=BetIndex, y=C.OTAReturns, color="1X2 A Returns")) +
  geom_line(aes(x=BetIndex, y=C.AHHReturns, color="AH H Returns")) +
  geom_line(aes(x=BetIndex, y=C.AHAReturns, color="AH A Returns")) +
  geom_line(aes(x=BetIndex, y=C.Returns, color="Total"),
            linetype="twodash") +
  scale_color_manual(name="Bet Type", values=c("1X2 H Returns" = "blue",
                                               "1X2 A Returns" = "coral",
                                               "AH H Returns"="green",
```

```
                                          "AH A Returns" = "purple",
                                          "Total"="red")) +
  theme_light() + coord_cartesian(xlim=c(0, 100), ylim=c(-40,20)) +
  geom_hline(yintercept=0, color="black", linetype="dotted", alpha=.5) +
  theme_light() +
  labs(title="Cum. Winnings (First 100 Games)\n using our Betting Model",
       caption="All matches 06-20", x="Match Index", y="Returns (unit)")

ggsave(path = "./writeup/images", filename = "model_01.png",
       plot=cr1, unit="cm", width=15, height=15)
ggsave(path = "./writeup/images", filename = "model_02.png",
       plot=cr2, unit="cm", width=15, height=15)

## ACCURACY OF THE MODEL ----
n.OTH <- nrow(matches[matches$OTHomeBet > 0,])
n.OTA <- nrow(matches[matches$OTAwayBet > 0,])
n.AHH <- nrow(matches[matches$AHHomeBet > 0,])
n.AHA <- nrow(matches[matches$AHAwayBet > 0,])
n.BetsPlaced <- n.OTH + n.OTA + n.AHH + n.AHA
n.MatchesBet <- nrow(matches[which(matches$OTHomeBet > 0 |
                                   matches$OTAwayBet > 0 |
                                   matches$AHHomeBet > 0 |
                                   matches$AHAwayBet > 0),])
n.BetsMatrix <- matrix(round(c(n.OTH, n.OTA, n.AHH, n.AHA,
    n.BetsPlaced), 5),
                       ncol = 5)
accuracy.ot.h <- nrow(matches[matches$OTHReturns > 0,]) / n.OTH * 100
accuracy.ot.a <- nrow(matches[matches$OTAReturns > 0,]) / n.OTA * 100
accuracy.ah.h <- nrow(matches[matches$AHHReturns > 0,]) / n.AHH * 100
accuracy.ah.a <- nrow(matches[matches$AHAReturns > 0,]) / n.AHA * 100
accuracy.overall <- 100 * (nrow(matches[matches$OTHReturns > 0,]) +
                       nrow(matches[matches$OTAReturns > 0,]) +
                       nrow(matches[matches$AHHReturns > 0,]) +
                       nrow(matches[matches$AHAReturns > 0,])) /
                           n.BetsPlaced

accuracy <- matrix(c(accuracy.ot.h, accuracy.ot.a, accuracy.ah.h,
                     accuracy.ah.a, accuracy.overall), ncol = 5)

total.winnings <- sum(matches$OTHReturns) + sum(matches$OTAReturns) +
  sum(matches$AHHReturns) + sum(matches$AHAReturns)

bet.winnings <- matrix(c(sum(matches$OTHReturns),
    sum(matches$OTAReturns),
                  sum(matches$AHHReturns), sum(matches$AHAReturns),
                  total.winnings), ncol = 5)

bet.analysis <- matrix(c(n.BetsMatrix, bet.winnings, accuracy), nrow = 5,
                       byrow = F, dimnames = list(c('1x2 H', '1x2 A', 'AH
                           H',
                                                    'AH A', 'Overall'),
                                                  c('Bets Placed','Winnings',
                                                    'Accuracy (%)')))
bet.analysis
```

```
271
272 ## IGNORING THE 'LOW' PERFORMING LEAGUES (ALTERNATIVE METHOD) ----
273 #From our analysis, whilst all leagues performed well, we noticed the
        German and
274 #French, and English/Scottish Level 2 Leagues.
275
276 #First, we copy the bets we placed earlier into new columns:
277 matches$OTHomeBet.alt <- with(matches, OTHomeBet)
278 matches$OTAwayBet.alt <- with(matches, OTAwayBet)
279 matches$AHHomeBet.alt <- with(matches, AHHomeBet)
280 matches$AHAwayBet.alt <- with(matches, AHAwayBet)
281 #Removing bets placed in France, Germany and Level 2, Eng/Sco:-
282 for (i in 1:N){
283   if (matches$Div[i] %in% c("D1", "F1", "E2", "E3", "SC1")){
284     matches$OTHomeBet.alt[i] <- 0; matches$OTAwayBet.alt[i] <- 0;
285     matches$AHHomeBet.alt[i] <- 0; matches$AHAwayBet.alt[i] <- 0
286   }
287 }
288
289 #And doing the same with returns:-
290 matches$OTHRet.alt <- with(matches, OTHReturns)
291 matches$OTARet.alt <- with(matches, OTAReturns)
292 matches$AHHRet.alt <- with(matches, AHHReturns)
293 matches$AHARet.alt <- with(matches, AHAReturns)
294 for (i in 1:N){
295   if (matches$Div[i] %in% c("D1", "F1", "E2", "E3", "SC1")){
296     matches$OTHRet.alt[i] <- 0; matches$OTARet.alt[i] <- 0;
297     matches$AHHRet.alt[i] <- 0; matches$AHARet.alt[i] <- 0
298   }
299 }
300 #Cumulative Returns
301 matches$C.OTHRet.alt <- with(matches,0); matches$C.OTARet.alt <-
        with(matches,0)
302 matches$C.AHHRet.alt <- with(matches,0); matches$C.AHARet.alt <-
        with(matches,0)
303 matches$C.Returns.alt <- with(matches, 0)
304
305 matches$C.OTHRet.alt[1] <- matches$OTHRet.alt[1]
306 matches$C.OTARet.alt[1] <- matches$OTARet.alt[1]
307 matches$C.AHHRet.alt[1] <- matches$AHHRet.alt[1]
308 matches$C.AHARet.alt[1] <- matches$AHARet.alt[1]
309 matches$C.Returns.alt[1] <- matches$OTHRet.alt[1] +
        matches$OTARet.alt[1] +
310   matches$AHHRet.alt[1] + matches$AHARet.alt[1]
311
312 for (i in 2:N){
313   matches$C.Returns.alt[i] <-
314     matches$C.Returns.alt[i-1]+matches$OTHRet.alt[i]+matches$OTARet.alt[i]+
315     matches$AHHRet.alt[i]+matches$AHARet.alt[i]}
316 for (i in 2:N){
317   matches$C.OTHRet.alt[i] <- matches$C.OTHRet.alt[i-1] +
          matches$OTHRet.alt[i]}
318 for (i in 2:N){
```

```
319   matches$C.OTARet.alt[i] <- matches$C.OTARet.alt[i-1] +
          matches$OTARet.alt[i]}
320 for (i in 2:N){
321   matches$C.AHHRet.alt[i] <- matches$C.AHHRet.alt[i-1] +
          matches$AHHRet.alt[i]}
322 for (i in 2:N){
323   matches$C.AHARet.alt[i] <- matches$C.AHARet.alt[i-1] +
          matches$AHARet.alt[i]}
324
325 #Accuracy
326 n.OTH.alt <- nrow(matches[matches$OTHomeBet.alt > 0,])
327 n.OTA.alt <- nrow(matches[matches$OTAwayBet.alt > 0,])
328 n.AHH.alt <- nrow(matches[matches$AHHomeBet.alt > 0,])
329 n.AHA.alt <- nrow(matches[matches$AHAwayBet.alt > 0,])
330 n.BetsP.alt <- n.OTH.alt + n.OTA.alt + n.AHH.alt + n.AHA.alt
331 n.BetsMatrix.alt <- matrix(round(c(n.OTH.alt, n.OTA.alt, n.AHH.alt,
        n.AHA.alt,
332                               n.BetsP.alt), 5), ncol = 5)
333
334 accalt.ot.h <- nrow(matches[matches$OTHRet.alt > 0,]) / n.OTH.alt * 100
335 accalt.ot.a <- nrow(matches[matches$OTARet.alt > 0,]) / n.OTA.alt * 100
336 accalt.ah.h <- nrow(matches[matches$AHHRet.alt > 0,]) / n.AHH.alt * 100
337 accalt.ah.a <- nrow(matches[matches$AHARet.alt > 0,]) / n.AHA.alt * 100
338 accalt.ovr <- 100 * (nrow(matches[matches$OTHRet.alt > 0,]) +
339                      nrow(matches[matches$OTARet.alt > 0,]) +
340                      nrow(matches[matches$AHHRet.alt > 0,]) +
341                      nrow(matches[matches$AHARet.alt > 0,])) /
                           n.BetsP.alt
342 accalt <- matrix(c(accalt.ot.h, accalt.ot.a, accalt.ah.h,
343                    accalt.ah.a, accalt.ovr), ncol = 5)
344
345 total.wins.alt <- sum(matches$OTHRet.alt) + sum(matches$OTARet.alt) +
346   sum(matches$AHHRet.alt) + sum(matches$AHARet.alt)
347
348 bet.wins.alt <- matrix(c(sum(matches$OTHRet.alt),
        sum(matches$OTARet.alt),
349                          sum(matches$AHHRet.alt), sum(matches$AHARet.alt),
350                          total.wins.alt), ncol = 5)
351
352 bet.analysis.alt <- matrix(c(n.BetsMatrix.alt, bet.wins.alt, accalt),
        nrow = 5,
353                          byrow = F, dimnames = list(c('1x2 H', '1x2 A', 'AH
                               H',
354                                                      'AH A', 'Overall'),
355                                                    c('Bets Placed','Winnings',
356                                                      'Accuracy (%)')))
357 bet.analysis.alt
358
359
360 ## COMPARISON OF ACCURACY AGAINST RANDOMLY PLACED BETS ----
361 #To compare our method against randomly placed bets, we will choose
362 #a random subset of N matches and find the winnings.
363
364 N <- nrow(matches)
```

```
365
366 for (i in 1:5){
367   #Reset
368   matches$randbet.oth <- with(matches, 0)
369   matches$randbet.ota <- with(matches, 0)
370   matches$randbet.ahh <- with(matches, 0)
371   matches$randbet.aha <- with(matches, 0)
372
373   #Finding Sample
374   sel.oth <- matches[sample(1:N, size = n.OTH, replace = F),]
375   sel.ota <- matches[sample(1:N, size = n.OTA, replace = F),]
376   sel.ahh <- matches[sample(1:N, size = n.AHH, replace = F),]
377   sel.aha <- matches[sample(1:N, size = n.AHA, replace = F),]
378
379   #Placing Bets
380   for (j in 1:nrow(matches[matches$OTHomeBet == 1,]))
381     {sel.oth$randbet.oth[j] <- 1}
382   for (j in (nrow(matches[matches$OTHomeBet == 1,])+1):
383       nrow(matches[matches$OTHomeBet == 2,])){sel.oth$randbet.oth[j] <-
          2}
384   for (j in (nrow(matches[matches$OTHomeBet == 2,])+1):
385       nrow(matches[matches$OTHomeBet == 3,])){sel.oth$randbet.oth[j] <-
          3}
386
387   for (j in 1:nrow(matches[matches$OTAwayBet == 1,]))
388     {sel.ota$randbet.ota[j] <- 1}
389   for (j in (nrow(matches[matches$OTAwayBet == 1,])+1):
390       nrow(matches[matches$OTAwayBet == 2,])){sel.ota$randbet.ota[j] <-
          2}
391   for (j in (nrow(matches[matches$OTAwayBet == 2,])+1):
392       nrow(matches[matches$OTAwayBet == 3,])){sel.ota$randbet.ota[j] <-
          3}
393
394   for (j in 1:nrow(matches[matches$AHHomeBet == 1,]))
395     {sel.ahh$randbet.ahh[j] <- 1}
396   for (j in (nrow(matches[matches$AHHomeBet == 1,])+1):
397       nrow(matches[matches$AHHomeBet == 2,])){sel.ahh$randbet.ahh[j] <-
          2}
398   for (j in (nrow(matches[matches$AHHomeBet == 2,])+1):
399       nrow(matches[matches$AHHomeBet == 3,])){sel.ahh$randbet.ahh[j] <-
          3}
400
401   for (j in 1:nrow(matches[matches$AHAwayBet == 1,]))
402     {sel.aha$randbet.aha[j] <- 1}
403   for (j in (nrow(matches[matches$AHAwayBet == 1,])+1):
404       nrow(matches[matches$AHAwayBet == 2,])){sel.aha$randbet.aha[j] <-
          2}
405   for (j in (nrow(matches[matches$AHAwayBet == 2,])+1):
406       nrow(matches[matches$AHAwayBet == 3,])){sel.aha$randbet.aha[j] <-
          3}
407
408   #Results
409   sel.oth$returns <- with(sel.oth, 0); sel.ota$returns <- with(sel.ota,
          0)
```

```
410   sel.ahh$returns <- with(sel.ahh, 0); sel.aha$returns <- with(sel.aha,
          0)
411
412   for (i in 1:n.OTH){
413     if (sel.oth$FTR[i] == "H"){
414       sel.oth$returns[i] <- (sel.oth$BbAvH[i]-1) *
              sel.oth$randbet.oth[i]}
415     else{sel.oth$returns[i] <- -sel.oth$randbet.oth[i]}}
416   for (i in 1:n.OTA){
417     if (sel.ota$FTR[i] == "A"){
418       sel.ota$returns[i] <- (sel.ota$BbAvA[i]-1) *
              sel.ota$randbet.ota[i]}
419     else{sel.ota$returns[i] <- -sel.ota$randbet.ota[i]}}
420   for (i in 1:n.AHH){
421     if (sel.ahh$ah.res[i] == "hm"){
422       sel.ahh$returns[i] <- (sel.ahh$BbAvAHH[i]-1) *
              sel.ahh$randbet.ahh[i]}
423     else if (sel.ahh$ah.res[i] == "hfhm"){
424       sel.ahh$returns[i] <-
              ((sel.ahh$BbAvAHH[i]-1)*0.5*sel.ahh$randbet.ahh[i])-
425         0.5*sel.ahh$randbet.ahh[i]}
426     else{sel.ahh$returns[i] <- -sel.ahh$randbet.ahh[i]}
427   }
428   for (i in 1:n.AHA){
429     if (sel.aha$ah.res[i] == "aw"){
430       sel.aha$returns[i] <- (sel.aha$BbAvAHA[i]-1) *
              sel.aha$randbet.aha[i]}
431     else if (sel.aha$ah.res[i] == "hfaw"){
432       sel.aha$returns[i] <-
              ((sel.aha$BbAvAHA[i]-1)*0.5*sel.aha$randbet.aha[i])-
433         0.5*sel.aha$randbet.aha[i]}
434     else{sel.aha$returns[i] <- -sel.aha$randbet.aha[i]}
435   }
436
437   #Assessing accuracy
438   acc.ran.oth <- nrow(sel.oth[sel.oth$returns > 0,]) / n.OTH * 100
439   acc.ran.ota <- nrow(sel.ota[sel.ota$returns > 0,]) / n.OTA * 100
440   acc.ran.ahh <- nrow(sel.ahh[sel.ahh$returns > 0,]) / n.AHH * 100
441   acc.ran.aha <- nrow(sel.aha[sel.aha$returns > 0,]) / n.AHA * 100
442
443   win.ran.oth <- sum(sel.oth$returns);win.ran.ota <- sum(sel.ota$returns)
444   win.ran.ahh <- sum(sel.ahh$returns);win.ran.aha <- sum(sel.aha$returns)
445
446   accmat <- matrix(c(n.OTH, n.OTA, n.AHH, n.AHA,
447                      win.ran.oth, win.ran.ota, win.ran.ahh, win.ran.aha,
448                      acc.ran.oth, acc.ran.ota, acc.ran.ahh, acc.ran.aha),
449                    nrow = 4, ncol = 3, byrow = F)
450   colnames(accmat) <- c('Bets', 'Wins', 'Acc')
451   rownames(accmat) <- c('1x2 H', '1x2 A', 'AH H', 'AH A')
452   print(accmat)
453 }
454
455 ## MULTIPLE RUNS OF RANDOM BET STRATEGY ----
```

```
#Here, we make bets with the same probability(placing bet) as with our
    model
#which may/may not result in different number of bets placed.

par(mfrow = c(2,2)); nRuns <- 30; alpha0 <- 1/nRuns #No. runs you wish
    to have. More = slower.

plot(matches$BetIndex, matches$C.OTHReturns, col = alpha("red"), type =
    'l',
    ylim = c(-2500, 10), xlab = 'Index', ylab = 'Returns', main = '1X2
        Home Win')
for (i in 1:nRuns){
  #Reset the random bet, 1X2 Home
  p1 <- nrow(matches[matches$OTHomeBet == 1,]) / nrow(matches[(n+1):N,])
  p2 <- nrow(matches[matches$OTHomeBet == 2,]) / nrow(matches[(n+1):N,])
  p3 <- nrow(matches[matches$OTHomeBet == 3,]) / nrow(matches[(n+1):N,])
  p0 <- 1 - (p1 + p2 + p3)

  matches$rand.Bet.OTH <- with(matches, 0)
  matches$rand.Ret.OTH <- with(matches, 0)
  matches$rand.Bet.OTH <- with(matches, rand.Bet.OTH +
                               rdiscrete(n = nrow(matches), values = 0:3,
                                         probs=c(p0, p1, p2, p3)))
  #For the random bets, we have 'placed bets' on the 05/06 season,
  #we will ignore it for the plots
  for(i in n:N){
    if (matches$FTR[i] == "H"){
      matches$rand.Ret.OTH[i] <- (matches$BbAvH[i]-1) *
          matches$rand.Bet.OTH[i]}
      else{matches$rand.Ret.OTH[i] <- -matches$rand.Bet.OTH[i]}}

  matches$rand.CumR.OTH <- with(matches, 0)
  matches$rand.CumR.OTH[n] <- matches$rand.Ret.OTH[n]
  for (i in n:N){matches$rand.CumR.OTH[i] <-
    matches$rand.CumR.OTH[i-1] + matches$rand.Ret.OTH[i]}

  lines(matches$BetIndex, matches$rand.CumR.OTH, col = alpha("blue",
      alpha0), type = 'l')
}
lines(matches$BetIndex, matches$C.OTHReturns, col = alpha("red"), type =
    'l')
lines(matches$BetIndex, matches$C.OTHRet.alt, col = alpha("green"), type
    = 'l')

plot(matches$BetIndex, matches$C.OTAReturns, col = alpha("red"), type =
    'l',
    ylim = c(-3500, 10), xlab = 'Index', ylab = 'Returns', main = '1X2
        Away Win')
for (i in 1:nRuns){
  p1 <- nrow(matches[matches$OTAwayBet == 1,]) / nrow(matches[(n+1):N,])
  p2 <- nrow(matches[matches$OTAwayBet == 2,]) / nrow(matches[(n+1):N,])
  p3 <- nrow(matches[matches$OTAwayBet == 3,]) / nrow(matches[(n+1):N,])
  p0 <- 1 - (p1 + p2 + p3)

```

```
500   #Reset the random bet, 1X2 Away
501   matches$rand.Bet.OTA <- with(matches, 0)
502   matches$rand.Ret.OTA <- with(matches, 0)
503   matches$rand.Bet.OTA <- with(matches, rand.Bet.OTA +
504                                rdiscrete(n = nrow(matches), values = 0:3,
505                                          probs=c(p0, p1, p2, p3)))
506   for (i in n:N){
507     if (matches$FTR[i] == "A"){
508       matches$rand.Ret.OTA[i] <- (matches$BbAvA[i]-1) *
          matches$rand.Bet.OTA[i]}
509     else{matches$rand.Ret.OTA[i] <- -matches$rand.Bet.OTA[i]}}
510
511   matches$rand.CumR.OTA <- with(matches, 0)
512   matches$rand.CumR.OTA[n] <- matches$rand.Ret.OTA[n]
513   for (i in n:N){matches$rand.CumR.OTA[i] <-
514     matches$rand.CumR.OTA[i-1] + matches$rand.Ret.OTA[i]}
515
516   lines(matches$BetIndex,matches$rand.CumR.OTA,col=alpha("blue",alpha0),type='l')
517 }
518 lines(matches$BetIndex, matches$C.OTAReturns, col = alpha("red"), type =
      'l')
519 lines(matches$BetIndex, matches$C.OTARet.alt, col = alpha("green"), type
      = 'l')
520
521 plot(matches$BetIndex, matches$C.AHHReturns, col = alpha("red"), type =
      'l',
522      ylim = c(-3500, 10), xlab = 'Index', ylab = 'Returns', main = 'AH
         Home Win')
523 for (i in 1:nRuns){
524   p1 <- nrow(matches[matches$AHHomeBet == 1,]) / nrow(matches[(n+1):N,])
525   p2 <- nrow(matches[matches$AHHomeBet == 1,]) / nrow(matches[(n+1):N,])
526   p3 <- nrow(matches[matches$AHHomeBet == 1,]) / nrow(matches[(n+1):N,])
527   p0 <- 1 - (p1 + p2 + p3)
528   #Reset the random bet, AH Home
529   matches$rand.Bet.AHH <- with(matches, 0)
530   matches$rand.Ret.AHH <- with(matches, 0)
531   matches$rand.Bet.AHH <- with(matches, rand.Bet.AHH +
532                                rdiscrete(n = nrow(matches), values = 0:3,
533                                          probs=c(p0, p1, p2, p3)))
534   for (i in n:N){
535     if (matches$ah.res[i] == "hm"){
536       matches$rand.Ret.AHH[i] <- (matches$BbAvAHH[i]-1) *
          matches$rand.Bet.AHH[i]}
537     else if (matches$ah.res[i] == "hfhm"){
538       matches$rand.Ret.AHH[i] <-
          (matches$BbAvAHH[i]-1)*0.5*matches$rand.Bet.AHH[i]-(0.5*matches$rand.Bet.AHH[i])}
539     else{matches$rand.Ret.AHH[i] <- -matches$rand.Bet.AHH[i]}}
540
541   matches$rand.CumR.AHH <- with(matches, 0)
542   matches$rand.CumR.AHH[n] <- matches$rand.Ret.AHH[n]
543   for (i in n:N){matches$rand.CumR.AHH[i] <-
544     matches$rand.CumR.AHH[i-1] + matches$rand.Ret.AHH[i]}
545
546   lines(matches$BetIndex,matches$rand.CumR.AHH,col=alpha("blue",alpha0),type='l')
```

```
}
lines(matches$BetIndex, matches$C.AHHReturns, col = alpha("red"), type = 'l')
lines(matches$BetIndex, matches$C.AHHRet.alt, col = alpha("green"), type = 'l')

plot(matches$BetIndex, matches$C.AHAReturns, col = alpha("red"), type = 'l',
     ylim = c(-3500, 10), xlab = 'Index', ylab = 'Returns', main = 'AH Away Win')
for (i in 1:nRuns){
  p1 <- nrow(matches[matches$AHAwayBet == 1,]) / nrow(matches[(n+1):N,])
  p2 <- nrow(matches[matches$AHAwayBet == 2,]) / nrow(matches[(n+1):N,])
  p3 <- nrow(matches[matches$AHAwayBet == 3,]) / nrow(matches[(n+1):N,])
  p0 <- 1 - (p1 + p2 + p3)
  #Reset the random bet, AH Away
  matches$rand.Bet.AHA <- with(matches, 0)
  matches$rand.Ret.AHA <- with(matches, 0)
  matches$rand.Bet.AHA <- with(matches, rand.Bet.AHA +
                               rdiscrete(n = nrow(matches), values = 0:3,
                                         probs=c(p0, p1, p2, p3)))
  for (i in n:N){
    if (matches$ah.res[i] == "aw"){
      matches$rand.Ret.AHA[i] <- (matches$BbAvAHA[i]-1) *
          matches$rand.Bet.AHA[i]}
    else if (matches$ah.res[i] == "hfaw"){
      matches$rand.Ret.AHA[i] <-
          (matches$BbAvAHA[i]-1)*0.5*matches$rand.Bet.AHA[i]-(0.5*matches$rand.Bet.AHA[i])}
    else{matches$rand.Ret.AHA[i] <- -matches$rand.Bet.AHA[i]}}

  matches$rand.CumR.AHA <- with(matches, 0)
  matches$rand.CumR.AHA[n] <- matches$rand.Ret.AHA[n]
  for (i in n:N){matches$rand.CumR.AHA[i] <-
    matches$rand.CumR.AHA[i-1] + matches$rand.Ret.AHA[i]}

  lines(matches$BetIndex,matches$rand.CumR.AHA,col=alpha("blue",alpha0),type='l')
}
lines(matches$BetIndex, matches$C.AHAReturns, col = alpha("red"), type = 'l')
lines(matches$BetIndex, matches$C.AHARet.alt, col = alpha("green"), type = 'l')

#We assess the accuracy of the *last* run only (this is easiest to code and
#allows us to run it without a large number of outputs
#For the n.Ran.___, we add that BetIndex > 0 to avoid the 05/06 season.
n.Ran.OTH <- nrow(matches[matches$rand.Bet.OTH > 0 & matches$BetIndex > 0,])
n.Ran.OTA <- nrow(matches[matches$rand.Bet.OTA > 0 & matches$BetIndex > 0,])
n.Ran.AHH <- nrow(matches[matches$rand.Bet.AHH > 0 & matches$BetIndex > 0,])
n.Ran.AHA <- nrow(matches[matches$rand.Bet.AHA > 0 & matches$BetIndex > 0,])
```

```
588 n.RanBets <- n.Ran.OTH + n.Ran.OTA + n.Ran.AHH + n.Ran.AHA
589 n.RanMatx <- matrix(c(n.Ran.OTH, n.Ran.OTA, n.Ran.AHH, n.Ran.AHA,
        n.RanBets),
590                     ncol = 5)
591
592 acc.Ran.OTH <- nrow(matches[matches$rand.Ret.OTH > 0,]) / n.Ran.OTH * 100
593 acc.Ran.OTA <- nrow(matches[matches$rand.Ret.OTA > 0,]) / n.Ran.OTA * 100
594 acc.Ran.AHH <- nrow(matches[matches$rand.Ret.AHH > 0,]) / n.Ran.AHH * 100
595 acc.Ran.AHA <- nrow(matches[matches$rand.Ret.AHA > 0,]) / n.Ran.AHA * 100
596 acc.Ran.ovr <- 100 * (nrow(matches[matches$rand.Ret.OTH > 0,]) +
597                       nrow(matches[matches$rand.Ret.OTA > 0,]) +
598                       nrow(matches[matches$rand.Ret.AHH > 0,]) +
599                       nrow(matches[matches$rand.Ret.AHA > 0,])) /
                              n.RanBets
600
601 acc.Ran <- matrix(c(acc.Ran.OTH, acc.Ran.OTA, acc.Ran.AHH, acc.Ran.AHA,
602                     acc.Ran.ovr), ncol = 5)
603
604 ran.winnings <- sum(matches$rand.Ret.OTH) + sum(matches$rand.Ret.OTA) +
605   sum(matches$rand.Ret.AHH) + sum(matches$rand.Ret.AHA)
606
607 ran.winningsMtx <-
        matrix(c(sum(matches$rand.Ret.OTH),sum(matches$rand.Ret.OTA),
608                        sum(matches$rand.Ret.AHH),
                               sum(matches$rand.Ret.AHA),
609                        ran.winnings), ncol = 5)
610
611 bet.analysis.random <- matrix(c(n.RanMatx, ran.winningsMtx, acc.Ran),
        nrow = 5,
612                       byrow = F, dimnames = list(c('1x2 H', '1x2 A', 'AH
                          H',
613                                                  'AH A', 'Overall'),
614                                                c('Bets Placed','Winnings',
615                                                  'Accuracy (%)')))
616 bet.analysis.random
```

Appendix E

Project Diary

Meeting One — 25/09/20

Maha, Alun and I discussed the start of my project: how it will be assessed, what I need to do each week, ETC. We spoke about where the data can be found (`football-data.co.uk`), how to import it, how to use it and how it is formatted. We went over the basics of probabilities and their relationships with odds and the bookmaker's commissions (typically 5%). Before next week, I will read the Kaunitz et al paper and look at replicating the simple steps (means, standard deviations) with my data.

Meeting Two — 02/10/20

We reviewed the Kaunitz paper, and looked at my R code used to complete the first steps. Before next week, I need to find out how `football-data.co.uk` source their data; do exploratory work on R (such as histograms and other basic plots) for either the 1X2 market across multiple seasons and leagues, or look at goals data for one league.

Meeting Three — 09/10/20

After Meeting Two, I made plots of observed v. bookmaker probabilities using fixed points (this ended up with some very small bins): we discussed how it would be better to make sure the bin sizes are equal, instead. Maha spoke about the need for ensuring my code is well-commented, and to start properly writing up what graphs show (it will make my final write up easier). Finally, we realised `football-data.co.uk` renamed their `BbAvH` column to `AvgH`: I had come across problems with using data from multiple seasons.

Meeting Four — 16/10/20

We looked into problems I faced with previous code, caused by small sample sizes. The point was made that `tapply` can be used to find the mean of a bin, rather than taking the midpoint, creating far more accurate plots and linear

models. We discussed methods for comparing accuracy over time (statistics such as R^2, mean squared error, ETC. or tests such as Kendall's Tau or Pearson's Rank), and spoke about the affect of *competitive balance*, and how we can quantify it to use in part of our analysis.

Meeting Five — 23/10/20

With Alun, we spoke about what plots I have created showed, and how I can use them in my project as a point for discussion. We discussed why RMSE is a better measure of accuracy than mean square error, median square error or absolute square error. Alun also showed me some extra papers, with methods and statistics (P_1 and P_2) that can be applied to the project. We discussed the future of the project, and how I will continue to progress. Finally, we spoke about the need to record everything I do, and to include more significant figures in my values for R^2.

Meeting Six — 06/11/20

Up until now, I had imported the `football-data.co.uk` datasets individually: Maha showed me a much better method using `for` loops and the `paste0` R command, which cut down a 1,000+ line document to around 100. We spoke about a number of papers I found about competitive balance, and the coefficients they used to quantify it. We spoke about using a binomial regression model, using $1/O_i$ (inverse of odds) as a predictor, and adding the season and leagues as additional ones, to see what affect they have. Finally, we spoke about what information would be included in my exploratory data analysis section.

Meeting Seven — 13/11/20

We spoke about using an ordinal logistic model, rather than binomial, and the pros and cons of each (OLM is simpler to code and run, but we may have multicollinearity between Home Win and Away Win probabilities). Alun demonstrated how the `ordinal` package worked, with the `clm` and `clmm` commands. We discussed whether or not the country would be treated as a fixed factor, and finally, we spoke about looking into how often the bookmakers 'favourite' won, and different ways we could define this.

Meeting Eight — 20/11/20

Maha gave me information about the literature review, and how I could prepare for this: for each paper I use, write down a summary of the paper, including:

- what the paper is about,
- how the researchers found the results (their method),
- the results they showed,
- and similarities and differences with my paper.

Maha recommended I create a table with these headings to fill in each time I use a new paper. We also discussed my ethical approval (the first check is on December 7th 2020).

Meeting Nine — 27/11/20

Due to coursework deadlines, I hadn't done much new work on my project, so we went over my progress so far, discussion what I've showed so far. We concluded I'm making good headway into the project, and how I can improve my results. We discussed other packages and techniques I can use, such as the `erer`, `marginal` and `mass` packages.

Meeting Ten — 16/12/20

Throughout the last week, I'd been working through a number of papers and summarising them (as per Maha's suggestion, Meeting Eight). We discussed what I need to show by the January check. I'd also made a number of plots using the `ggplot2` package, which Alun and I went over, making a number of comments about them. We looked at changing the parameters for the density plot (default is Gaussian kernal estimate, and $n = 512$). We discussed these plots at length and how the matches with a high probability of a draw are likely due to match fixing or when a draw suits both teams. At length, we discussed the difference in density plots for each league, and why some leagues appeared to have a unimodal distribution, and others having a trimodal distribution. Alun suggested using Odds Portal to check any unusual odds. We also discussed a tile plot I created, and discussed that grouping 5+ goals was key: small bin sizes are not helpful. I also gave a brief outline of possible sections for my first section, which we modified, and what progress I'm expected to make over Christmas. Finally, we discussed the Under/Over 2.5 Goals market, and the Asian Handicap market, and how I could look into the accuracy of these over the Christmas break, and throughout Semester Two.

Meeting Eleven — 18/01/21

Code and Content

Over the Christmas break, I'd made a section of code to find R^2, RMSE, P_1, and P_2 for each league that was 205 lines long: we discussed how a `for` loop would be much better, like how we did to read in the data. For the tile plot, we discussed possible ways of including bin size: we both agreed inserting a table below would be clearest. Maha confirmed a question I had about including code in-text, rather than referring to it in the appendices.

Write-Up

In my submitted document, I couldn't get references to work: Maha recommended the `natbib` LaTeXpackage, and we spoke about how APA and Harvard are the best referencing systems. Maha also made it clear figures should be in

the middle of the text, where they are needed in context, so the reader doesn't have to continually refer back and forth. A small conclusion at the end of each chapter is required, as well as a final concluding chapter with our findings and discussion. Figure labels need to be below the figure; table labels above. Maha also sent an annotated copy of my write-up.

Meeting Twelve — 28/01/21

As well as making changes based on our previous meeting, I'd found a good paper discussing different leagues styles of play: as a result, I conducted principal components analysis with three new variables: `predAcc` (predictive accuracy), `imbalance` (a measure of the competitive imbalance in a league), and `attack` (a measure based on the shots per goal: more shots implies a more attacking league). The Kruskal test is a good addition, as well as potentially running PCA on the seasons, rather than leagues. Finally, after Maha's comments on my draft about normalising the probabilities, I found Shin's Method of Normalisation, which I will apply and compare over the next week.

Meeting Thirteen — 16/02/21

In the two weeks, I finalised my *elite* data analysis, and moved onto the English and Scottish leagues, including the Under/Over 2.5 Goals and Asian Handicap markets. Due to Maha being off ill, I caught Alun up with my progress, to which he made a few comments: I should look into the leverage of my models, especially with the Under/Over 2.5 Goals market; the χ^2 test for independence between goals is good, but due to some expected values being low, I would need to group goals together: it also doesn't take into account the order of goals, so perhaps a different test (Spearman or Pearson) is better. Alun also introduced me to the Skellam distribution, based on the Poisson distribution, as a way to model goals, and recommended I read about bivariate Poisson distributions.

Meeting Fourteen — 02/03/21

I made large amounts of progress in the two weeks, and started a new section, creating a betting model based on bookmaker's odds. Alun suggested making a few changes, such as changing the range of my bets for different markets and changing the units I bet on. We discussed ignoring the lower performing leagues and markets (such as German Bundesliga and Level 2 in the English and Scottish leagues, and the U/O market).

Meeting Fifteen — 16/03/21

Alun and I went through all of my outputs (figures, tables, ETC.), and discussed what each one showed. He recommended I talk about the amount of coding knowledge I have acquired throughout the dissertation, and challenges I have faced throughout the course of the project, and how I have overcame them. We

also discussed the presentation, and how I can both prepare for it, and what to include in it.

Meeting Sixteen — 30/03/21

Alun and I discussed the final stages of the project. We went through my presentation slides, suggesting possible areas of improvement, and what needs to be included in my script. We went through what will be completed by the second check deadline, and what I will submit then, as well as a discussion into whether or not it is necessary to include the correlation tests in the report (Kendall's Tau, Spearman Rank, and Goodman-Kruskal-Gamma tests): adding something with no value can make the report harder to read. Finally, we spoke about the conclusion chapter, namely the inclusion of common themes and contradictory findings in it.

Appendix F

Word Count

Chapter	*Prose*	*Figures*	*Tables*	**Total**
Abstract	0	0	0	0
1: Introduction	2047	0	173	2220
2: *Elite*	5436	99	488	6023
3: English/Scottish	3454	235	375	4064
4: Proposed Method	58	0	0	58
5: Conclusion	40	0	0	40
Total	11109	334	1036	**12365**

Appendix G

Ethical Approval Certificate

Certificate of Ethical Approval

Applicant:	Joseph Pym
Project Title:	Assessing the Accuracy of Betting Odds in Football

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Medium Risk

Date of approval:	14 Dec 2020
Project Reference Number:	P114620