Northeastern University
College of Professional Studies

Percepta Project: Final Report

Sharon Appoline Rosary, Naveenkumar Govindasamy, David Joseph Johnson
ALY6015-21591 SEC 10 Intermediate Analytics
Instructor: Vladimir Shapiro
Date:19 February, 2022

Introduction

The Campeonato Nacional de Liga de Primera División commonly known as La Liga is the top professional soccer division for men in the Spanish football league system. It consists of 20 teams; the bottom three teams are relegated to the Segunda División and replaced by the top two teams as well as the play-off winner from the Segunda Division.

Over the years, 62 teams have competed in La Liga, winning nine titles. Real Madrid has won the championship a record 34 times. La Liga has been the top league in Europe in each of the seven years from 2013 to 2019 (calculated using accumulated figures from five preceding seasons) and has led Europe for 22 of the 60 ranked years up to 2019, more than any other country.

The dataset consists of data collected from each match including data points like:

- Full-time result
- Half-time result
- Home Team Goals
- Goals
- Shots
- Shots on Target
- Red cards
- Yellow cards

Betting information pertaining to return values of 6 companies corresponding to Home Team Win, Away Team Win and Draw is also included in the dataset.

The goal of the analysis is to find relevant insights on predicting the winning team and Betting returns. Feature Selection, Correlation of Different Factors, Fitting and enhancing the performance of Predictive Models will be executed while answering the business questions.

Overview of Business Questions and Methods Used:

- Predict the winning team (Home team or away team).
  From the preliminary analysis, we found that teams are more likely to win on their home ground. A predictive model to determine the winning team whilst including other factors

can be built using logistic regression technique.

- What are the factors influencing the returns of betting companies?
  It is a general assumption that the returns from betting companies vary w.r.t the match statistics. Using multiple regression analysis, we can find the factors that influence the returns of betting companies. This method facilitates the analysis of relationships between single dependent variable (betting return rates) and multiple independent variables (match statistics such as goals, wins, red cards and yellow cards, etc.)
- Do all betting companies give the same returns?
  While placing bets it is confusing to choose a company that yields maximum results, hence differences in mean returns of companies is analyzed. ANOVA test can be used to statistically validate if all companies provide similar returns.

Data Cleaning

```
# Loading Dataset
raw_data <- read.csv("/Users/davidjjohn97/Northeastern/Quarter_2/Term_3/Intermediate_Analytics/La_Lig
a_2019.csv",header = TRUE)
# Datasets from 2015-2019 are combined to give a larger training set for predictive model
raw_data_extra <- read.csv("/Users/davidjjohn97/Northeastern/Quarter_2/Term_3/Intermediate_Analytics/
La_Liga_2015-19.csv",header = TRUE)
# Data Cleaning
#Remove empty rows and columns.
laliga <- remove_empty(raw_data,which = c("rows","cols"),quiet=TRUE)
laliga_extra <- remove_empty(raw_data_extra,which = c("rows","cols"),quiet=TRUE)
names(laliga)
```

```
##   [1] "Div"         "Date"        "Time"        "HomeTeam"    "AwayTeam"
##   [6] "FTHG"        "FTAG"        "FTR"         "HTHG"        "HTAG"
##  [11] "HTR"         "HS"          "AS"          "HST"         "AST"
##  [16] "HF"          "AF"          "HC"          "AC"          "HY"
##  [21] "AY"          "HR"          "AR"          "B365H"       "B365D"
##  [26] "B365A"       "BWH"         "BWD"         "BWA"         "IWH"
##  [31] "IWD"         "IWA"         "PSH"         "PSD"         "PSA"
##  [36] "WHH"         "WHD"         "WHA"         "VCH"         "VCD"
##  [41] "VCA"         "MaxH"        "MaxD"        "MaxA"        "AvgH"
##  [46] "AvgD"        "AvgA"        "B365.2.5"    "B365.2.5.1"  "P.2.5"
##  [51] "P.2.5.1"     "Max.2.5"     "Max.2.5.1"   "Avg.2.5"     "Avg.2.5.1"
##  [56] "AHh"         "B365AHH"     "B365AHA"     "PAHH"        "PAHA"
##  [61] "MaxAHH"      "MaxAHA"      "AvgAHH"      "AvgAHA"      "B365CH"
##  [66] "B365CD"      "B365CA"      "BWCH"        "BWCD"        "BWCA"
##  [71] "IWCH"        "IWCD"        "IWCA"        "PSCH"        "PSCD"
##  [76] "PSCA"        "WHCH"        "WHCD"        "WHCA"        "VCCH"
##  [81] "VCCD"        "VCCA"        "MaxCH"       "MaxCD"       "MaxCA"
##  [86] "AvgCH"       "AvgCD"       "AvgCA"       "B365C.2.5"   "B365C.2.5.1"
##  [91] "PC.2.5"      "PC.2.5.1"    "MaxC.2.5"    "MaxC.2.5.1"  "AvgC.2.5"
##  [96] "AvgC.2.5.1"  "AHCh"        "B365CAHH"    "B365CAHA"    "PCAHH"
## [101] "PCAHA"       "MaxCAHH"     "MaxCAHA"     "AvgCAHH"     "AvgCAHA"
```

```
#Drop unwanted variables.
laliga_new <- laliga[-c(2:3,42:105)]
laliga_extra <- laliga_extra[c(5:22)]
names(laliga_new)
```

```
##  [1] "Div"      "HomeTeam" "AwayTeam" "FTHG"     "FTAG"     "FTR"
##  [7] "HTHG"     "HTAG"     "HTR"      "HS"       "AS"       "HST"
## [13] "AST"      "HF"       "AF"       "HC"       "AC"       "HY"
## [19] "AY"       "HR"       "AR"       "B365H"    "B365D"    "B365A"
## [25] "BWH"      "BWD"      "BWA"      "IWH"      "IWD"      "IWA"
## [31] "PSH"      "PSD"      "PSA"      "WHH"      "WHD"      "WHA"
## [37] "VCH"      "VCD"      "VCA"
```

```
#Check for duplicate records.
get_dupes(laliga_new)
```

```
## No variable names specified - using all columns.
```

```
## No duplicate combinations found of: Div, HomeTeam, AwayTeam, FTHG, FTAG, FTR, HTHG, HTAG, HTR, ...
and 30 other variables
```

```
##  [1] Div        HomeTeam   AwayTeam   FTHG       FTAG       FTR
##  [7] HTHG       HTAG       HTR        HS         AS         HST
## [13] AST        HF         AF         HC         AC         HY
## [19] AY         HR         AR         B365H      B365D      B365A
## [25] BWH        BWD        BWA        IWH        IWD        IWA
## [31] PSH        PSD        PSA        WHH        WHD        WHA
## [37] VCH        VCD        VCA        dupe_count
## <0 rows> (or 0-length row.names)
```

```
get_dupes(laliga_extra)
```

```
## No variable names specified - using all columns.
```

```
## No duplicate combinations found of: FTHG, FTAG, FTR, HTHG, HTAG, HTR, HS, AS, HST, ... and 9 other
variables
```

```
##  [1] FTHG       FTAG       FTR        HTHG       HTAG       HTR
##  [7] HS         AS         HST        AST        HF         AF
## [13] HC         AC         HY         AY         HR         AR
## [19] dupe_count
## <0 rows> (or 0-length row.names)
```

```
#Change column names to meaningful ones.
names(laliga_new)[4:21] = c("fult_home_goal","fult_away_goal","fult_result","hlft_home_goal","hlft_away_goal","hlft_result","home_shot","away_shot","home_shot_on_target","away_shot_on_target","home_fouls","away_fouls","home_corners","away_corners","home_yellow","away_yellow","home_red","away_red")
names(laliga_extra) = c("fult_home_goal","fult_away_goal","fult_result","hlft_home_goal","hlft_away_goal","hlft_result","home_shot","away_shot","home_shot_on_target","away_shot_on_target","home_fouls","away_fouls","home_corners","away_corners","home_yellow","away_yellow","home_red","away_red")
names(laliga_new)
```

```
##  [1] "Div"                "HomeTeam"           "AwayTeam"
##  [4] "fult_home_goal"     "fult_away_goal"     "fult_result"
##  [7] "hlft_home_goal"     "hlft_away_goal"     "hlft_result"
## [10] "home_shot"          "away_shot"          "home_shot_on_target"
## [13] "away_shot_on_target" "home_fouls"         "away_fouls"
## [16] "home_corners"       "away_corners"       "home_yellow"
## [19] "away_yellow"        "home_red"           "away_red"
## [22] "B365H"              "B365D"              "B365A"
## [25] "BWH"                "BWD"                "BWA"
## [28] "IWH"                "IWD"                "IWA"
## [31] "PSH"                "PSD"                "PSA"
## [34] "WHH"                "WHD"                "WHA"
## [37] "VCH"                "VCD"                "VCA"
```

```
#Check for NAs
knitr::kable(sum(is.na(laliga_new)),col.names = "No. of NAs")
```

No. of NAs
0

Analysis

Exploratory Data Analysis

```
knitr::kable(psych::describe(laliga_new,na.rm = TRUE,skew = FALSE),caption = "Descriptive Statistics")
```

Descriptive Statistics

	vars	n	mean	sd	min	max	range	se
Div*	1	180	1.0000000	0.0000000	1.00	1.00	0.00	0.0000000
HomeTeam*	2	180	10.3722222	5.7285237	1.00	20.00	19.00	0.4269789
AwayTeam*	3	180	10.6277778	5.8328970	1.00	20.00	19.00	0.4347585
fult_home_goal	4	180	1.5555556	1.3084266	0.00	5.00	5.00	0.0975244

	vars	n	mean	sd	min	max	range	se
fult_away_goal	5	180	1.0500000	0.9874916	0.00	4.00	4.00	0.0736033
fult_result*	6	180	2.2000000	0.8282714	1.00	3.00	2.00	0.0617357
hlft_home_goal	7	180	0.6500000	0.8682803	0.00	4.00	4.00	0.0647178
hlft_away_goal	8	180	0.4277778	0.6166277	0.00	3.00	3.00	0.0459607
hlft_result*	9	180	2.1277778	0.7093199	1.00	3.00	2.00	0.0528696
home_shot	10	180	12.9555556	5.0404201	4.00	25.00	21.00	0.3756907
away_shot	11	180	10.3555556	4.1475876	2.00	24.00	22.00	0.3091429
home_shot_on_target	12	180	4.5555556	2.6787436	0.00	17.00	17.00	0.1996618
away_shot_on_target	13	180	3.4611111	1.8860378	0.00	11.00	11.00	0.1405770
home_fouls	14	180	13.2277778	4.2908641	4.00	28.00	24.00	0.3198221
away_fouls	15	180	13.9277778	3.9762290	5.00	28.00	23.00	0.2963706
home_corners	16	180	5.3722222	2.6384153	0.00	14.00	14.00	0.1966559
away_corners	17	180	4.3111111	2.2430249	0.00	11.00	11.00	0.1671852
home_yellow	18	180	2.4333333	1.5064109	0.00	6.00	6.00	0.1122812
away_yellow	19	180	2.7944444	1.5596061	0.00	8.00	8.00	0.1162462
home_red	20	180	0.1166667	0.3703388	0.00	2.00	2.00	0.0276034
away_red	21	180	0.1111111	0.3324009	0.00	2.00	2.00	0.0247757
B365H	22	180	2.5871111	1.4898440	1.12	10.00	8.88	0.1110464
B365D	23	180	3.8200000	1.0886946	2.87	9.50	6.63	0.0811465
B365A	24	180	4.4790000	3.1011559	1.30	21.00	19.70	0.2311465
BWH	25	180	2.5808333	1.4469439	1.12	10.00	8.88	0.1078488
BWD	26	180	3.8102778	1.0799997	2.85	9.50	6.65	0.0804984
BWA	27	180	4.4386111	2.9829633	1.30	20.00	18.70	0.2223370
IWH	28	180	2.5798889	1.4114844	1.13	9.60	8.47	0.1052058
IWD	29	180	3.7769444	1.0015960	2.85	8.80	5.95	0.0746546
IWA	30	180	4.3899444	2.9814594	1.30	19.50	18.20	0.2222249

	vars	n	mean	sd	min	max	range	se
PSH	31	180	2.6262778	1.4591702	1.12	9.99	8.87	0.1087601
PSD	32	180	3.9058333	1.1394068	2.96	9.95	6.99	0.0849264
PSA	33	180	4.6817778	3.4880263	1.33	25.50	24.17	0.2599821
WHH	34	180	2.5905000	1.4526463	1.11	9.50	8.39	0.1082739
WHD	35	180	3.7932778	1.0238348	2.90	9.00	6.10	0.0763121
WHA	36	180	4.5877222	3.4000368	1.32	23.00	21.68	0.2534238
VCH	37	180	2.5658333	1.4329053	1.10	10.00	8.90	0.1068025
VCD	38	180	3.8436667	1.0780999	2.90	9.50	6.60	0.0803568
VCA	39	180	4.4875556	3.3600199	1.30	26.00	24.70	0.2504411

Table 1: Descriptive Statistics of all numerical variables in LaLiga dataset

From the above descriptive statistics table, we can infer the following:

- There are 180 observations with 39 variables.
- The highest number of goals scored by the home team in a match is 5 whereas the maximum number of goals scored by the away team is 4 in a match.
- The maximum number of fouls made in a match is 28. The average fouls made by the home team and the away team are nearly equal.
- The maximum number of yellow cards issued in a match is 8 whereas the maximum number of red cards issued in a match is 2.
- The highest average betting returns among betting companies for the home team is 2.626 by Poker stars betting company whereas the highest betting returns for the away team is 4.58 by William hill (WH) betting company.
- The minimum return for home team among betting companies is 1.30 by Venture Capital (VC) betting company and the maximum return for the home team is 10 by Betting365, B.W and Venture Capital (VC) betting companies.
- The minimum return and maximum return among betting companies for the away team is 1.10 and 26 by Venture Capital (VC) betting company.

Top ten teams with maximum number of goals throughout the league:

The top ten teams who scored maximum number of goals in home ground throughout the league can be plotted using a bar plot.

```
#aggregate function to calculate total goals scored as a home team.
agg_home <- aggregate(laliga_new$fult_home_goal, by=list(Category=laliga_new$HomeTeam), FUN=sum)
home_data <- agg_home[order(- agg_home$x),]

#Bar plot to show the top ten teams w.r.t home goals
ggplot(head(home_data,10),aes(x=x,y=Category))+ggtitle("Top 10 teams w.r.t Home goals")+labs(x="No. o
f goals",y=" ")+ geom_bar(stat="identity", fill="slateblue1")+geom_text(mapping=aes(label=x),position
=position_dodge(width=0.9),cex=3,hjust=-0.1)
```

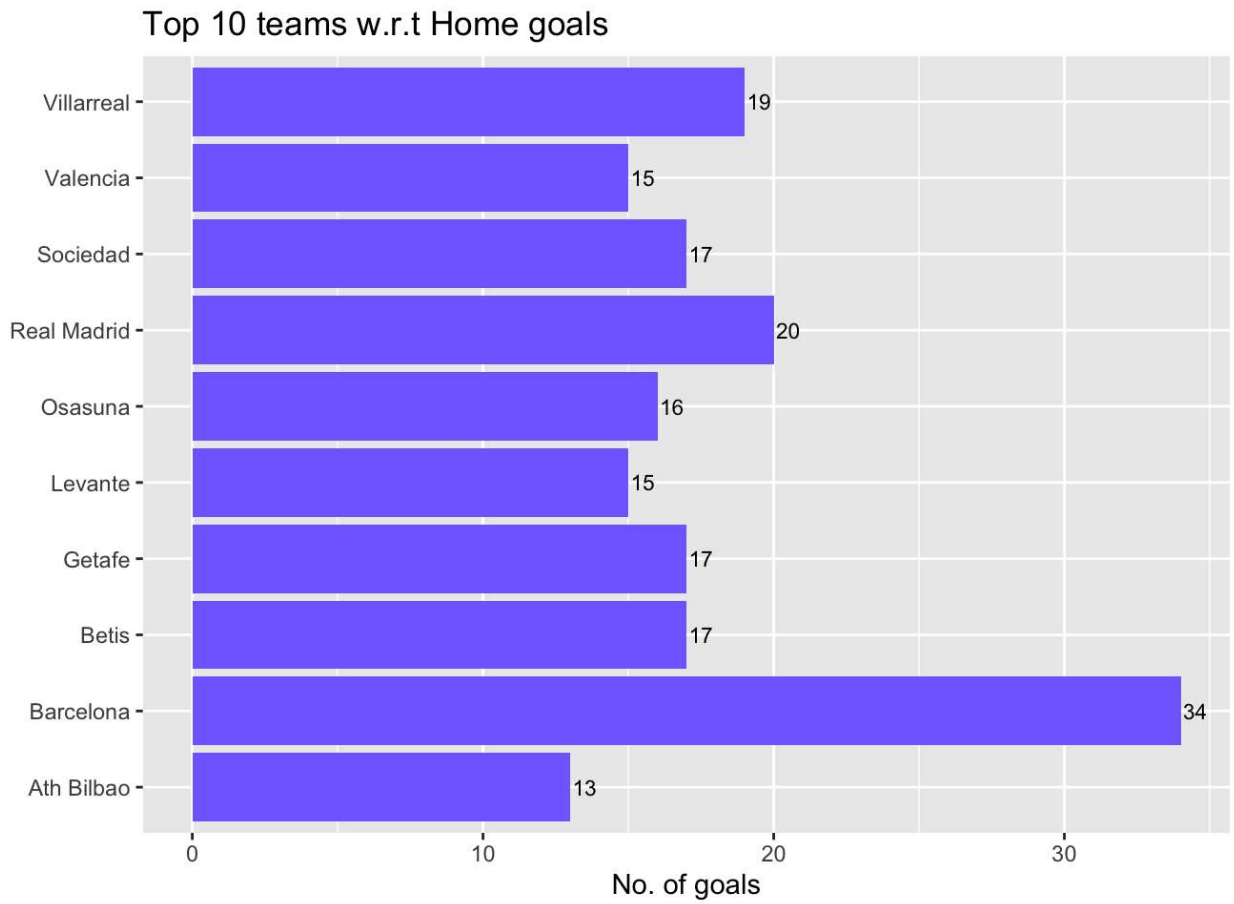

Figure 1: Bar-plot to depict the top 10 teams with highest home-goals

The above bar chart depicts that **Barcelona scored the maximum number of home goals** i.e., 34 ranking first in our top ten teams' table. The second highest team with the maximum number of home goals is Real Madrid i.e., 20.

A similar bar plot to depict the top ten teams who scored maximum number of goals in away ground throughout the league can be plotted.

```
#aggregate function to calculate total goals scored as a away team.
agg_away <- aggregate(laliga_new$fult_away_goal, by=list(Category=laliga_new$HomeTeam), FUN=sum)
away_data <- agg_away[order(- agg_away$x),]

#Bar plot to show the top ten teams w.r.t away goals
ggplot(head(away_data,10),aes(x=x,y=Category))+ggtitle("Top 10 teams w.r.t Away goals")+labs(x="No. o
f goals",y=" ")+ geom_bar(stat="identity", fill="slateblue1")+geom_text(mapping=aes(label=x),position
=position_dodge(width=0.9),cex=3,hjust=-0.1)
```

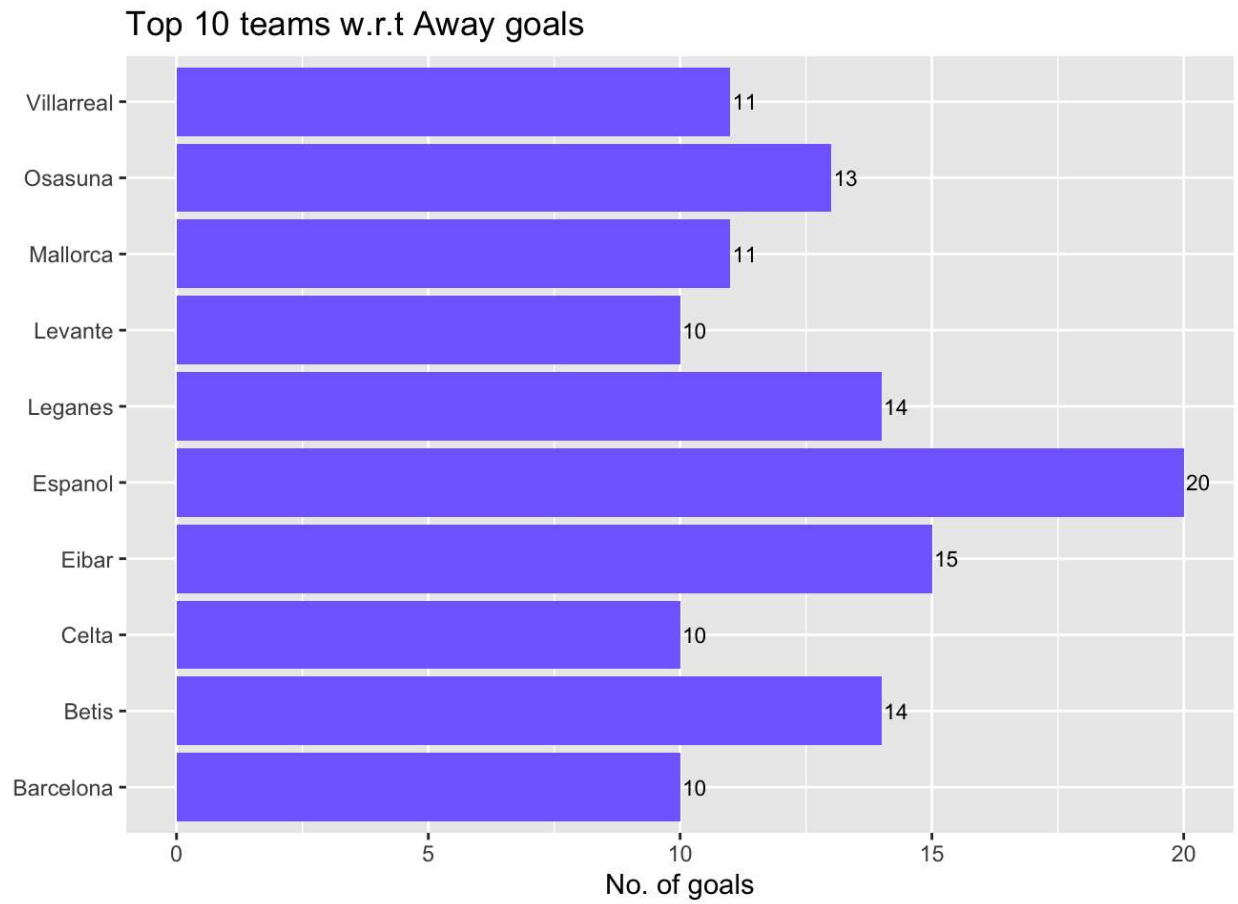

Figure 2: Bar-plot to depict the top 10 teams with highest away-goals

The above bar plot depicts that **Espanol scored the maximum number of away goals** i.e., 20 ranking first in our top ten team's table. The second highest team with the maximum number of away goals is Eibar i.e., 15. We can also observe that most of the teams appear in both the tables i.e., Teams like Villarreal, Osasuna, Betis, Barcelona have scored a fairly good number of goals on both home and away ground.

Top five teams with most red cards

The top teams with maximum number of red cards throughout the league can be calculated using the aggregate() function like below.

```
#aggregate function to calculate total red cards by home team.
agg_home_red <- aggregate(laliga_new$home_red, by=list(Category=laliga_new$HomeTeam), FUN=sum)
agg_home_red <- head(agg_home_red[order(- agg_home_red$x),],5)
rownames(agg_home_red) = c("1 ", "2 ", " 3", "4 ", " 5")
knitr::kable(agg_home_red[1:2],col.names = c("Team","No. of Red cards"),caption="Teams with most red
 cards (Home)")
```

Teams with most red cards (Home)

	Team	No. of Red cards
1	Betis	3
2	Celta	3
3	Levante	3
4	Barcelona	2
5	Osasuna	2

Table 2: Top 5 teams with highest Red cards on Home ground

Betis, Celta and Levante have been issued three red cards in their home ground throughout the league

```
#aggregate function to calculate total red cards by away team.
agg_away_red <- aggregate(laliga_new$away_red, by=list(Category=laliga_new$HomeTeam), FUN=sum)
agg_away_red <- head(agg_away_red[order(- agg_away_red$x),],5)
rownames(agg_away_red) = c("1 ", "2 ", " 3", "4 ", " 5")
knitr::kable(agg_away_red[1:2],col.names = c("Team","No. of Red cards"),caption="Teams with most red
 cards (Away)")
```

Teams with most red cards (Away)

	Team	No. of Red cards
1	Granada	3
2	Ath Madrid	2
3	Celta	2
4	Mallorca	2
5	Sociedad	2

Table 3: Top 5 teams with highest Red cards on Away ground

Granada has been issued three red cards in away ground throughout the league. We can see that Celta have been issued more red cards in both home ground and away ground.

Correlation between home goals and returns of different betting companies

The relationship between home goals and returns of different betting companies can be visually presented using a scatter plot.

```
ggplot(data=laliga_new,aes(B365H+BWH+IWH+PSH+WHH+VCH,fult_home_goal)) +
                        geom_point(color="slateblue1") +
ggtitle("Correlation between home goals and returns of different betting companies") + labs(y="Hom
            e goals",x= "Returns of betting companies (B365,BW,IW,PS,WH,VC)")
```

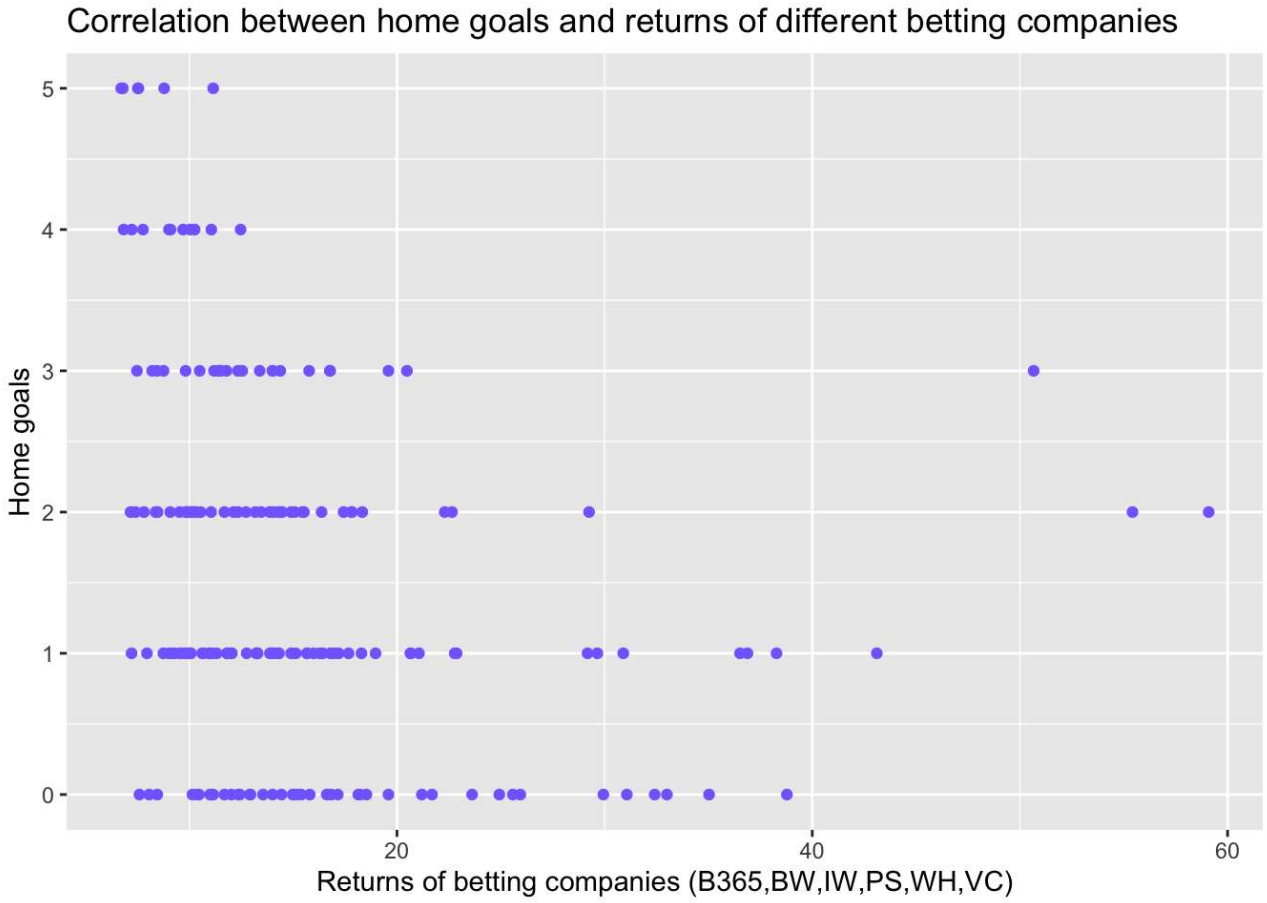

Figure 3: Scatter-plot to depict the relationship between Home-goals and return rates of different companies

The data points line up well, a horizontal line amongst these dots, and the line would clearly be a good fit to the data. However, the fact that the line would be horizontal means that the input values (that is, returns of betting companies) are irrelevant to the output values (that is, home goals). Any change in home goals, the returns are always going to be right around the same value.

```
x <- ggplot(data=laliga_new,aes(B365H+B365A+B365D,BWH+BWA+BWD)) +
                                    geom_point() +
  ggtitle("Correlation between Bet365 and B.W betting companies") +
                                  geom_smooth(method='lm')
y <- x + theme(plot.title = element_text(color="black", size=8, face="bold.italic"))
                                            y
```

```
## `geom_smooth()` using formula 'y ~ x'
```

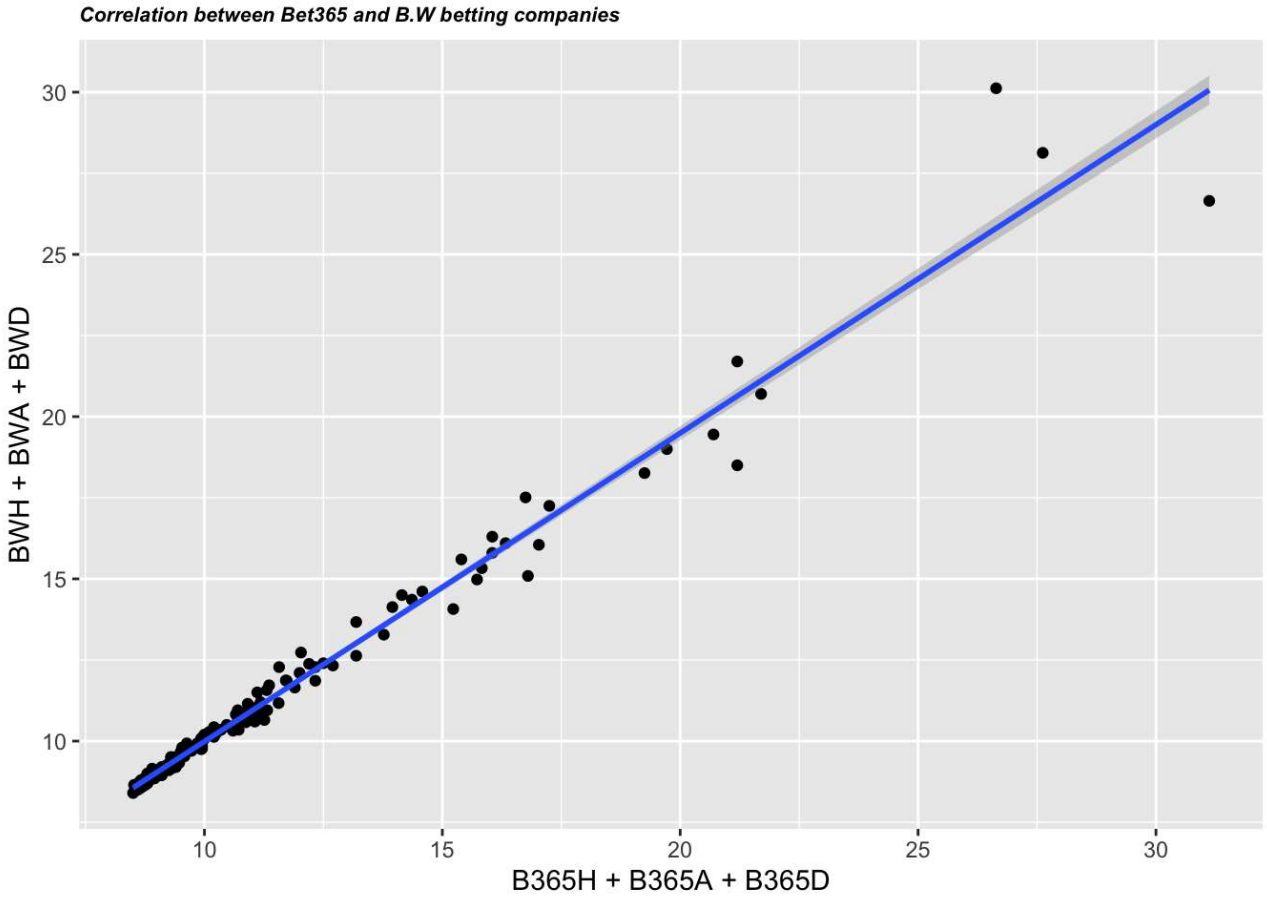

Figure 4: Scatter-plot to depict the relationship between Bet365 and B.W betting companies

```
w <- ggplot(data=laliga_new,aes(IWH+IWA+IWD,PSH+PSA+PSD)) +
                                    geom_point() +
  ggtitle("Correlation between Incarnate and Poker stars betting companies") +
                                  geom_smooth(method='lm')
z <- w + theme(plot.title = element_text(color="black", size=8, face="bold.italic"))
                                            z
```

```
## `geom_smooth()` using formula 'y ~ x'
```

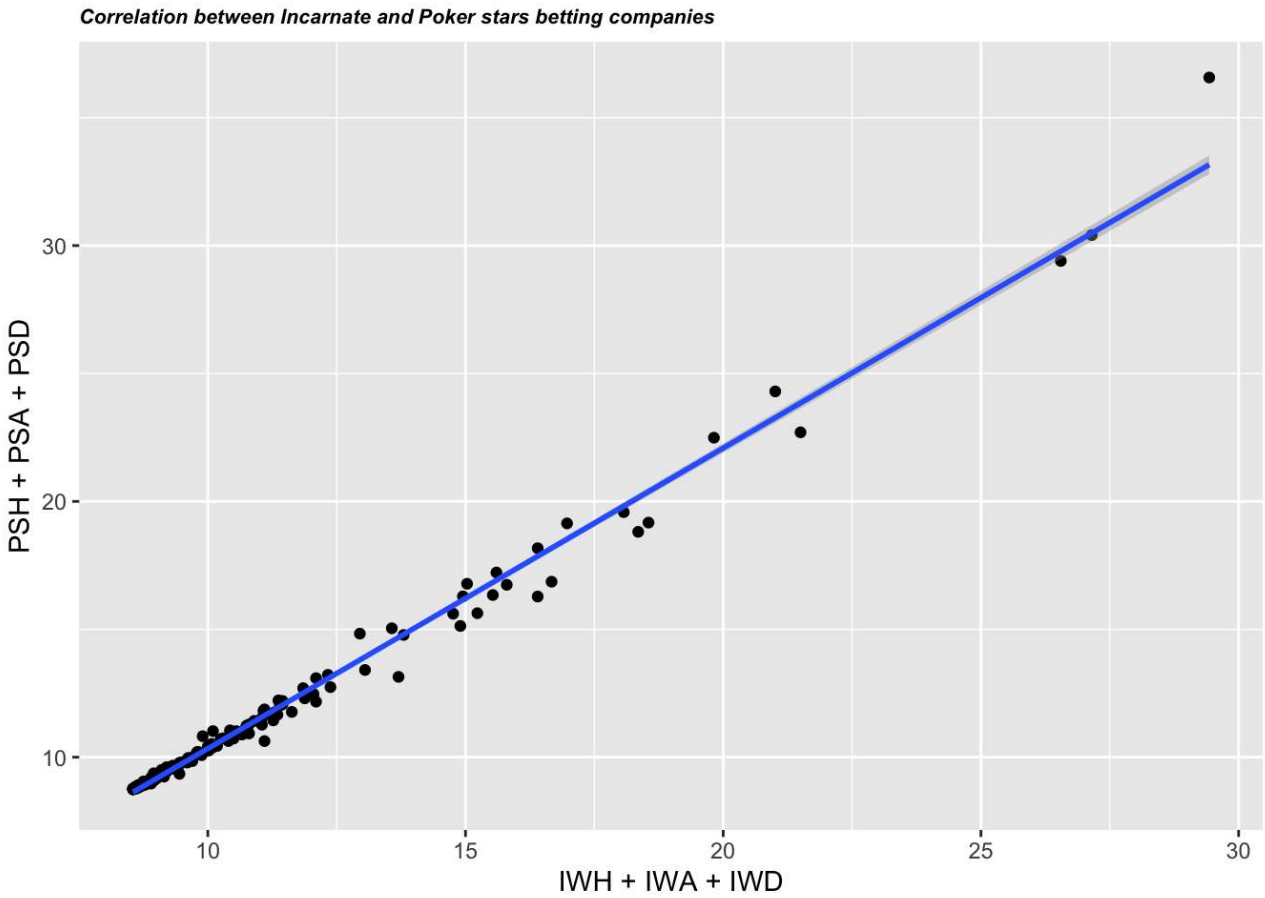

Figure 5: Scatter-plot to depict the relationship Incarnate and Poker stars betting companies

There is a strong positive correlation between the returns of the betting companies. i.e., As the returns of Betting365 company increase, the returns of B.W betting company also increase and vice versa.

Correlation between Shots vs shots on target by team

The correlation between shots and shots on target can be given as,

```
df <-laliga_new
dfe <- data.frame("Team", 0,0)
for (row in 1:nrow(df)) {
    ftr <- df[row, "fult_result"]
    date  <- df[row, "date"]

    if(ftr == "H") {
      dfe[nrow(dfe) + 1,] = c("Home", as.numeric(df[row, "home_shot"]),as.numeric(df[row, "home_shot_
on_target"]))
    }
    if(ftr == "A") {
      dfe[nrow(dfe) + 1,] = c("Away", as.numeric(df[row, "away_shot"]),as.numeric(df[row, "away_shot_
on_target"]))
    }
}
dfe <- dfe[-c(1), ]
colnames(dfe) <- c('Teams','Shots','Shots_on_Target')
dfe$Teams <- as.factor(dfe$Teams)
dfe$Shots <- as.numeric(dfe$Shots)
dfe$Shots_on_Target <- as.numeric(dfe$Shots_on_Target)

sct <- ggplot(dfe, aes(x=Shots, y=Shots_on_Target, color=Teams)) + geom_point(size=1) + geom_smooth(m
ethod=lm) + xlim(0,30) + ylim(0,20) +
  labs(x="Shots Attempted", y = "Shots on Target")+ ggtitle("Shots Vs Shots on Target by Team")
```

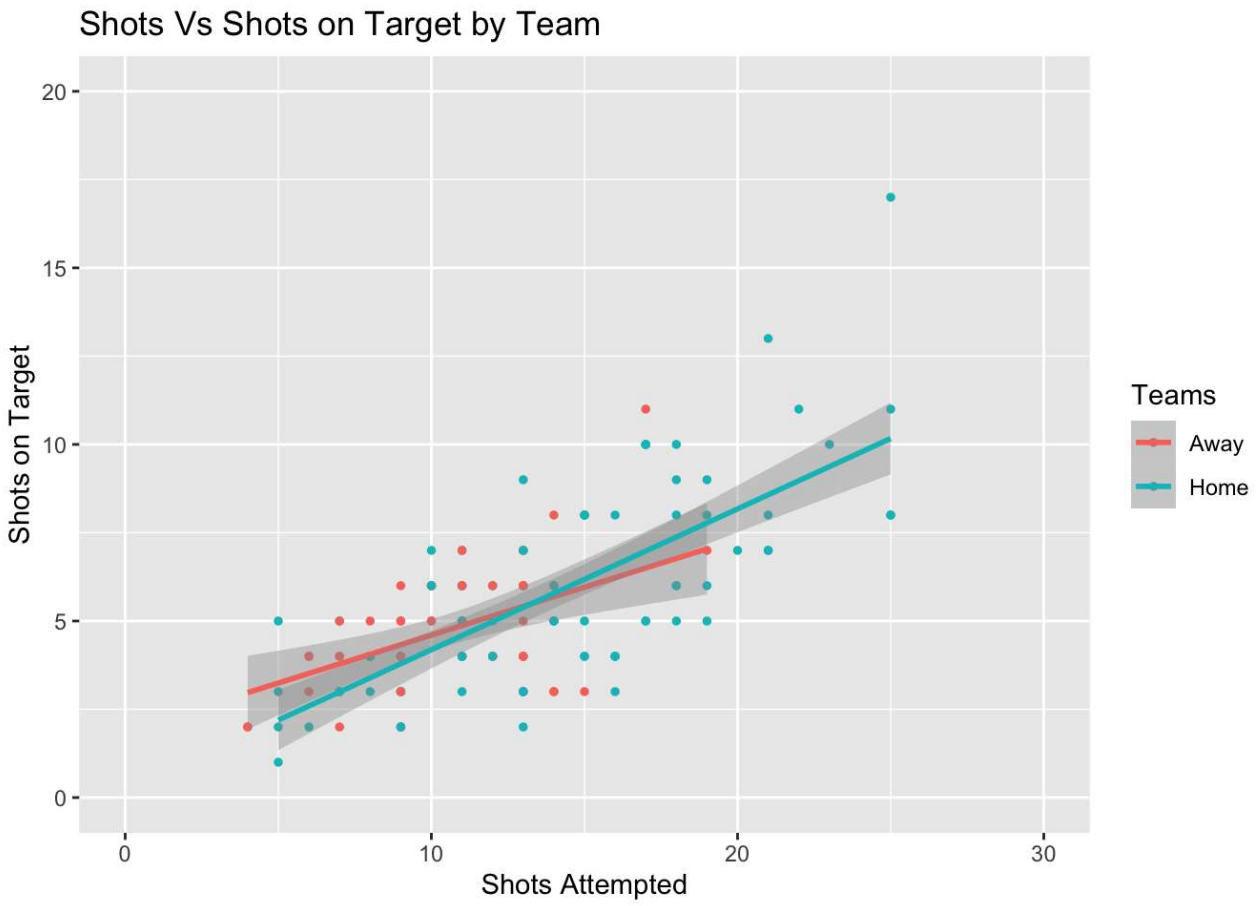

Figure 6: Scatter-plot to depict the relationship between Shots Vs Shots on Target by Team

The scatter plot above depicts the relationship between shots attempted and shots on target when the winning teams are playing on their home ground or away ground. We can observe that the teams play better on their home ground. This could be due to multiple factors like support of fans, familiarity with their home ground, etc.

Histogram of full-time results:

```
t <-as.data.frame(table(df$fult_result))
hist_win <-ggplot(df, aes(x=fult_result)) +
  geom_bar(fill="lightblue") +
  theme(axis.text.x = element_text(face="bold", color="black",
                           size=10),
          axis.text.y = element_text(face="bold", color="black",
                           size=10)) + ylim(0, 100) +
  labs(x="Winning Team", y = "Total Count") + scale_x_discrete(labels=c("A" = "Away Team", "D" = "Dra
w",
                              "H" = "Home Team")) +
  theme(plot.caption = element_text(hjust = 0.5)) +
  geom_text(aes(label = ..count..), stat = "count", vjust = 1.5, colour = "black") + ggtitle("Full Ti
me Results")
```

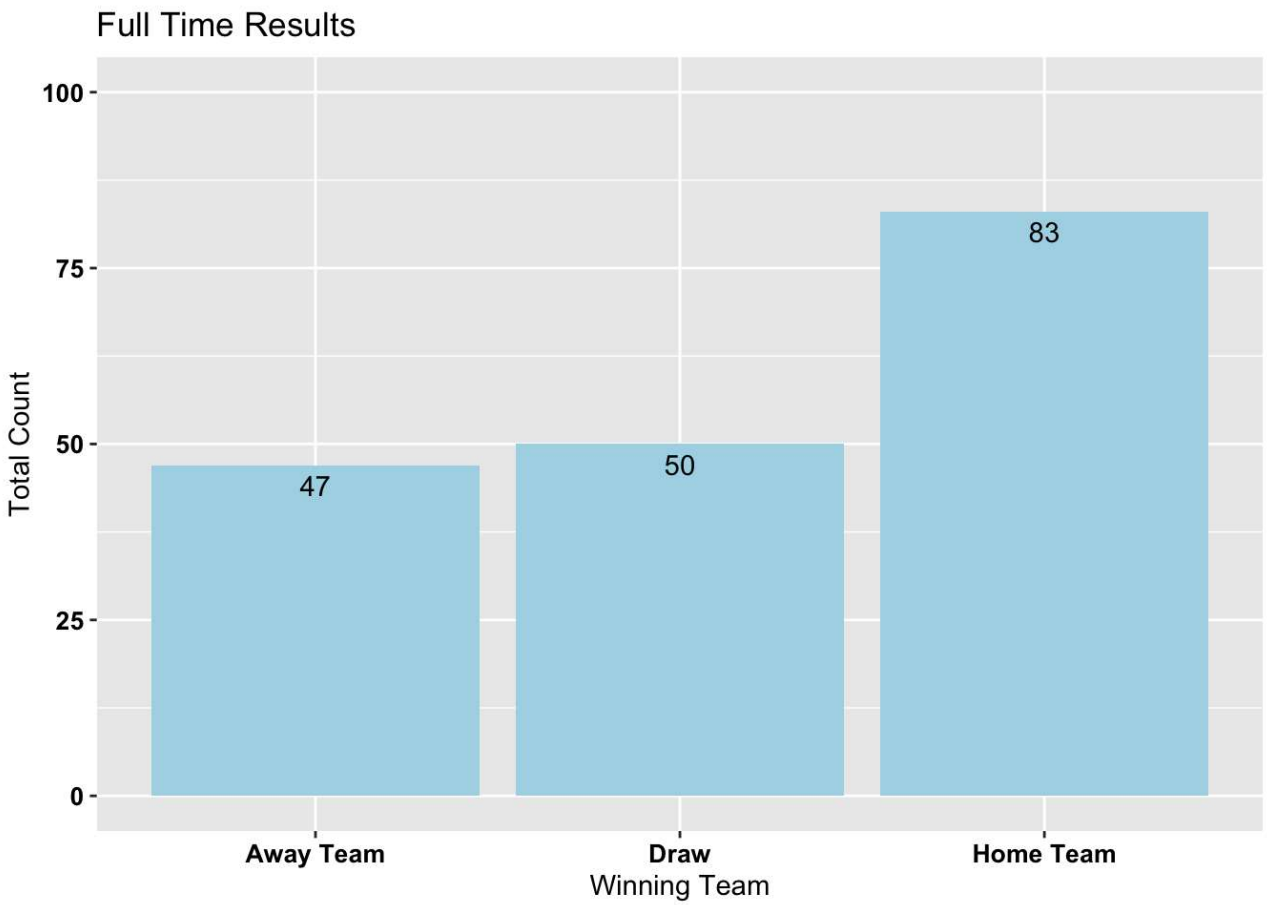

Figure 7: Histogram of Full Time Results

The plot above depicts the histogram of frequency of wins for away team, home team or draw. We can conclude that the total number of wins for home team is more (total of 83 wins) on the home ground.

Goal 1: Predict the winning team (Home team or away team)

Logistic Regression Models to Predict Winning Team

A predictive model to determine the winning team whilst including other factors can be built using logistic regression technique.

Logistic regression is a process of modelling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no and in this case Win/Loss.

```
# Logistic Regression Models to Predict Winning Team

# Preparing Data frame
laliga <- laliga_extra

## Drop unwanted variables.
names(laliga)
```

```
##  [1] "fult_home_goal"      "fult_away_goal"      "fult_result"
##  [4] "hlft_home_goal"      "hlft_away_goal"      "hlft_result"
##  [7] "home_shot"           "away_shot"           "home_shot_on_target"
## [10] "away_shot_on_target" "home_fouls"          "away_fouls"
## [13] "home_corners"        "away_corners"        "home_yellow"
## [16] "away_yellow"         "home_red"            "away_red"
```

```
head(laliga)
```

```
##   fult_home_goal fult_away_goal fult_result hlft_home_goal hlft_away_goal
## 1              1              0           H              0              0
## 2              1              3           A              0              1
## 3              1              1           D              0              0
## 4              2              1           H              1              0
## 5              0              1           A              0              0
## 6              4              4           D              1              1
##   hlft_result home_shot away_shot home_shot_on_target away_shot_on_target
## 1           D        11        11                   5                   2
## 2           A         7        17                   4                  11
## 3           D        14        12                   6                   3
## 4           H        16        11                   4                   5
## 5           D        13         4                   2                   2
## 6           D        12        14                   7                   7
##   home_fouls away_fouls home_corners away_corners home_yellow away_yellow
## 1         14          9            3            8           1           1
## 2         17         12            6            4           5           2
## 3         13         14            3            3           4           4
## 4         13         14            9            3           2           3
## 5         17         11            8            0           1           4
## 6         10         16            2            7           3           1
##   home_red away_red
## 1        0        0
## 2        0        1
## 3        1        0
## 4        0        0
## 5        1        0
## 6        0        0
```

```
## Converting attributes to numeric
laliga$hlft_home_goal <-as.numeric(laliga$hlft_home_goal)
laliga$hlft_away_goal <-as.numeric(laliga$hlft_away_goal)
laliga$home_shot <-as.numeric(laliga$home_shot)
laliga$away_shot <-as.numeric(laliga$away_shot)
laliga$home_fouls <-as.numeric(laliga$home_fouls)
laliga$away_fouls <-as.numeric(laliga$away_fouls)
laliga$home_corners <-as.numeric(laliga$home_corners)
laliga$away_corners <-as.numeric(laliga$away_corners)
laliga$home_yellow <-as.numeric(laliga$home_yellow)
laliga$away_yellow <-as.numeric(laliga$away_yellow)
laliga$home_red <-as.numeric(laliga$home_red)
laliga$away_red <-as.numeric(laliga$away_red)
laliga$home_shot_on_target <-as.numeric(laliga$home_shot_on_target)
laliga$away_shot_on_target <-as.numeric(laliga$away_shot_on_target)


# Remove Columns with Full Time Result => "Draw"
laliga_WL<-laliga[!(laliga$fult_result=="D"),]

# View Updated Data-Frame
head(laliga_WL)
```

```
##   fult_home_goal fult_away_goal fult_result hlft_home_goal hlft_away_goal
## 1              1              0           H              0              0
## 2              1              3           A              0              1
## 4              2              1           H              1              0
## 5              0              1           A              0              0
## 7              1              0           H              0              0
## 8              0              2           A              0              1
##   hlft_result home_shot away_shot home_shot_on_target away_shot_on_target
## 1           D        11        11                   5                   2
## 2           A         7        17                   4                  11
## 4           H        16        11                   4                   5
## 5           D        13         4                   2                   2
## 7           D         9        16                   2                   4
## 8           A         7        12                   2                   4
##   home_fouls away_fouls home_corners away_corners home_yellow away_yellow
## 1         14          9            3            8           1           1
## 2         17         12            6            4           5           2
## 4         13         14            9            3           2           3
## 5         17         11            8            0           1           4
## 7         18         15            2            9           2           1
## 8         11         17            8            4           2           2
##   home_red away_red
## 1        0        0
## 2        0        1
## 4        0        0
## 5        1        0
## 7        0        0
## 8        0        0
```

To fit the model, we must divide the data set into two parts: training and testing. The training data will be used to train the model, while the testing data will be used to test the model. This allows us to confirm that the model works effectively with new data and that we are not overfitting it.

The data is divided using the caret package's createDataPartition() function. I utilized a 70/30 split, which means that 70% of the observations will be used for training and 30% will be used for testing.

```
# Checking for Class Bias

## Check the proportion count.
pvt_prop<- table(laliga_WL$fult_result)
knitr::kable(pvt_prop,col.names = c("Level","Count"))
```

Level	Count
A	264
H	430

Table 4: Proportion of Wins for Home and Away Teams

```
## Barplot to show proportion of class in Full Time Result
barplot(pvt_prop, main="Barplot of Wins,to identify Class-Bias",col=rgb(0.8,0.1,0.1,0.6),width=0.1,space=0.5,ylim=c(0,500),xlab="Proportion")
```

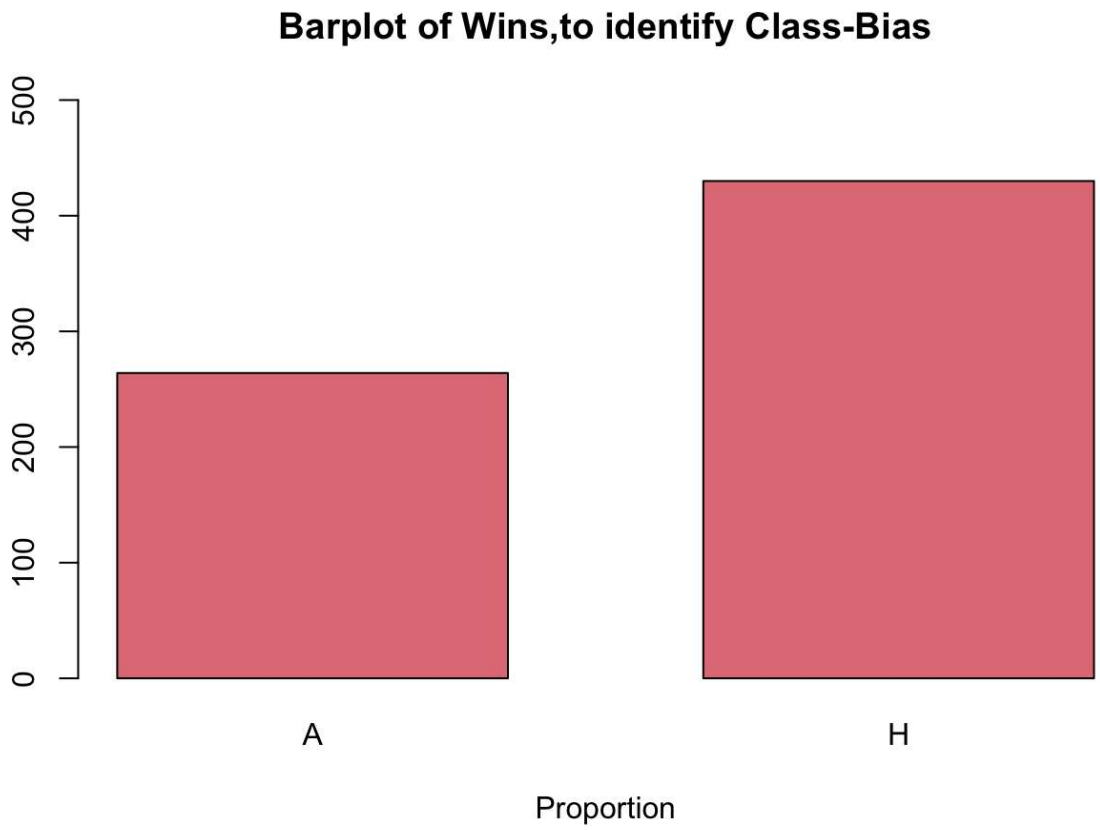

Figure 8: Bar-plot depicting Wins by Home and Away Teams

There is a class bias where the proportion of events is much larger than the proportion of non-events. So, we need to sample the observations in equal proportions to get better models.

To address the problem of class bias, we can draw wins from 0's and 1's for the training set in equal proportions. Then, we can put the rest of the data that is not included for training data (development sample) into test data (validation sample). Thus, the split can be done using sample() function.

```
# seed random generator for repeatability of samples.
set.seed(369)

# 70% - Training data and 30% - Testing data
laliga_WL$fult_result <- ifelse(laliga_WL$fult_result == "H", 1, 0)
train_fract <- 0.7
input_ones <- laliga_WL[which(laliga_WL$fult_result == 1),]
input_zeros <- laliga_WL[which(laliga_WL$fult_result == 0),]
no_bias_rw_no <- nrow(input_zeros)
input_zeros_no_bias <- input_zeros
input_ones_no_bias <- head(input_ones,no_bias_rw_no)

# Creation of training data:
rnd_row_index <-sample (1:nrow(input_zeros_no_bias),train_fract*no_bias_rw_no)
ones_training <- input_ones_no_bias[rnd_row_index,]
zeroes_training <- input_zeros_no_bias[rnd_row_index,]
train_set <- rbind(ones_training,zeroes_training)

# Creation of test data:
ones_test <- input_ones_no_bias[-rnd_row_index,]
zeroes_test <- input_zeros_no_bias[-rnd_row_index,]
test_set <- rbind(ones_test,zeroes_test)

# Check class bias
knitr::kable(table(train_set$fult_result),col.names = c("Level","Count"),caption="Training Data")
```

Training Data

Level	Count
0	184
1	184

Table 5: Proportion of Wins for Home and Away Teams (Training Data)

```
knitr::kable(table(test_set$fult_result),col.names = c("Level","Count"),caption="Test Data")
```

Test Data

Level	Count
0	80
1	80

Table 6: Proportion of Wins for Home and Away Teams (Test Data)

Initially, all the features are selected.

```
# Model1
win_predictive_model1 <- glm(fult_result ~ hlft_home_goal + hlft_away_goal + home_shot_on_target + aw
ay_shot_on_target + home_shot + away_shot + home_corners + away_corners + home_yellow + away_yellow +
home_red + away_red ,data = train_set,family = binomial(link = "logit"))
summary(win_predictive_model1)
```

```
##
## Call:
## glm(formula = fult_result ~ hlft_home_goal + hlft_away_goal +
##     home_shot_on_target + away_shot_on_target + home_shot + away_shot +
##     home_corners + away_corners + home_yellow + away_yellow +
##     home_red + away_red, family = binomial(link = "logit"), data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4922  -0.4586  -0.0014   0.4660   2.7262
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.412868   0.848467   0.487  0.62654
## hlft_home_goal        1.820601   0.321367   5.665 1.47e-08 ***
## hlft_away_goal       -1.514380   0.286005  -5.295 1.19e-07 ***
## home_shot_on_target   0.583269   0.111873   5.214 1.85e-07 ***
## away_shot_on_target  -0.667740   0.112485  -5.936 2.92e-09 ***
## home_shot            -0.111101   0.053052  -2.094  0.03624 *
## away_shot             0.118435   0.055431   2.137  0.03263 *
## home_corners         -0.005358   0.071969  -0.074  0.94066
## away_corners          0.034884   0.073971   0.472  0.63722
## home_yellow          -0.050815   0.107715  -0.472  0.63710
## away_yellow          -0.052725   0.120921  -0.436  0.66282
## home_red             -2.255732   0.717908  -3.142  0.00168 **
## away_red              0.675930   0.515688   1.311  0.18995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 510.16  on 367  degrees of freedom
## Residual deviance: 234.99  on 355  degrees of freedom
## AIC: 260.99
##
## Number of Fisher Scoring iterations: 6
```

Based on summary, a second model is created using only variables that proved to be significant.

```
# Based on summary a second model is created using only variables that proved to be significant
# Model2
win_predictive_model2 <- glm(fult_result ~ hlft_home_goal + hlft_away_goal + home_shot_on_target + aw
ay_shot_on_target + home_shot + away_shot + home_red,data = train_set,family = binomial(link = "logi
t"))
summary(win_predictive_model2)
```

```
##
## Call:
## glm(formula = fult_result ~ hlft_home_goal + hlft_away_goal +
##     home_shot_on_target + away_shot_on_target + home_shot + away_shot +
##     home_red, family = binomial(link = "logit"), data = train_set)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.54887  -0.45839  -0.00117   0.44278   2.63310
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.31896    0.68351   0.467   0.6408
## hlft_home_goal       1.79308    0.30742   5.833 5.46e-09 ***
## hlft_away_goal      -1.52479    0.28295  -5.389 7.09e-08 ***
## home_shot_on_target  0.58150    0.10874   5.348 8.90e-08 ***
## away_shot_on_target -0.66523    0.11190  -5.945 2.76e-09 ***
## home_shot           -0.11084    0.04533  -2.445   0.0145 *
## away_shot            0.12141    0.05250   2.313   0.0207 *
## home_red            -2.33594    0.70361  -3.320   0.0009 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 510.16  on 367  degrees of freedom
## Residual deviance: 237.52  on 360  degrees of freedom
## AIC: 253.52
##
## Number of Fisher Scoring iterations: 6
```

```
## Comparing regression coefficients (log-odds)
### Model 1
coef(win_predictive_model1)
```

```
##         (Intercept)      hlft_home_goal      hlft_away_goal home_shot_on_target
##         0.412867929         1.820601050        -1.514379868         0.583269281
## away_shot_on_target           home_shot           away_shot        home_corners
##        -0.667740101        -0.111100783         0.118434731        -0.005357581
##        away_corners         home_yellow         away_yellow            home_red
##         0.034884172        -0.050815432        -0.052724836        -2.255731765
##            away_red
##         0.675930393
```

```
### Model 2
coef(win_predictive_model2)
```

```
##         (Intercept)      hlft_home_goal      hlft_away_goal home_shot_on_target
##           0.3189554           1.7930810          -1.5247894           0.5815017
## away_shot_on_target           home_shot           away_shot            home_red
##          -0.6652307          -0.1108409           0.1214127          -2.3359393
```

```
## The two models can be compared using Akaike information criterion test
aic <- AIC(win_predictive_model1,win_predictive_model2)
aic
```

```
##                       df      AIC
## win_predictive_model1 13 260.9903
## win_predictive_model2  8 253.5216
```

Here I have fitted a logistic model as the predictor variable is categorical. I have fitted the model with all the variables of the dataset to understand the variables and their significance values. In the summary of the model, we see that the estimates of the coefficients are in log odds. We observed that hlft_home_goal, hlft_away_goa, away_shot_on_target, away_shot, home_corners, home_red are the most significant variables.

Since the ratio of data rows to the number of variables <40, we consider AIC values to determine which of multiple models is most likely to be the best model for a given dataset.

AIC test gives a AIC value of each model and the model with lower AIC value is preferred as it has less KL divergence and is therefore more suitable. Hence, Model 2 which has lower AIC (253.5216) value is considered a better model.

```
# log-odds
knitr::kable(coef(win_predictive_model2),col.names="log-odds")
```

	log-odds
(Intercept)	0.3189554
hlft_home_goal	1.7930810
hlft_away_goal	-1.5247894
home_shot_on_target	0.5815017
away_shot_on_target	-0.6652307
home_shot	-0.1108409
away_shot	0.1214127
home_red	-2.3359393

Table 7: Coefficients of Model2(log-odds)

```
# Odds
knitr::kable(exp(coef(win_predictive_model2)),col.names="Odds",caption="Conversion of log-odds to odd
s")
```

Conversion of log-odds to odds

	Odds
(Intercept)	1.3756900
hlft_home_goal	6.0079344
hlft_away_goal	0.2176669
home_shot_on_target	1.7887225
away_shot_on_target	0.5141549
home_shot	0.8950812
away_shot	1.1290907
home_red	0.0967196

Table 8: Coefficients of Model2(odds)

Interpretation:

- The odds of FullTimeResult(fult_result) being Home-Win, increase by a factor of 6.0079344 for each Half Time Home Goal(hlft_home_goal)
- The odds of FullTimeResult(fult_result) being Home-Win, decrease by a factor of 0.2176669 for each Half Time Away Goal(hlft_away_goal)
- The odds of FullTimeResult(fult_result) being Home-Win, increase by a factor of 0.5141549 for each home_shot_on_target
- The odds of FullTimeResult(fult_result) being Home-Win, decrease by a factor of 0.6652307 for each away_shot_on_target
- The odds of FullTimeResult(fult_result) being Home-Win, decrease by a factor of 0.1108409 for each home_shot
- The odds of FullTimeResult(fult_result) being Home-Win, decrease by a factor of 0.1214127 for each away_shot
- The odds of FullTimeResult(fult_result) being Home-Win, decrease by a factor of 2.3359393 for home_red

```
## Making predictions on test data
probabilities.test <- predict(win_predictive_model2, newdata = test_set, type = "response")
predicted.classes.min <- as.factor(ifelse(probabilities.test >= 0.5 ,"H","A"))

## Model Accuracy
test_set$fult_result <- as.factor(ifelse(as.numeric(test_set$fult_result) == 1, "H", "A"))
win_predictive_model2_CM <-confusionMatrix(predicted.classes.min,test_set$fult_result,positive="H")
win_predictive_model2_CM
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  A  H
##          A 65 11
##          H 15 69
##
##                Accuracy : 0.8375
##                  95% CI : (0.771, 0.891)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.675
##
##  Mcnemar's Test P-Value : 0.5563
##
##             Sensitivity : 0.8625
##             Specificity : 0.8125
##          Pos Pred Value : 0.8214
##          Neg Pred Value : 0.8553
##              Prevalence : 0.5000
##          Detection Rate : 0.4313
##    Detection Prevalence : 0.5250
##       Balanced Accuracy : 0.8375
##
##        'Positive' Class : H
##
```

The above confusion matrix depicts:

- This classification model is 83.75% accurate and it has a confidence interval of (0.771, 0.891) at 95% confidence level.
- The true positive value is 69 whereas the true negative value is 65. The false positive value is 15 whereas the false negative is 11. In this case, false positive(Predicting ‘Away Team Win’ but actually ‘Home Team Win’) miscalculation is more damaging for the analysis.

Plot the receiver operator characteristic curve

The receiver operator characteristic curve can be plotted using roc function from pROC package. It takes the response variable values and predicts probabilities as parameters.

```
# Plot ROC
ROC1 <- roc(test_set$fult_result,probabilities.test)
```

```
## Setting levels: control = A, case = H
```

```
## Setting direction: controls < cases
```

```
plot(ROC1,col="blue",ylab="Sensitivity - TP Rate", xlab="Specificity - FP Rate")
```

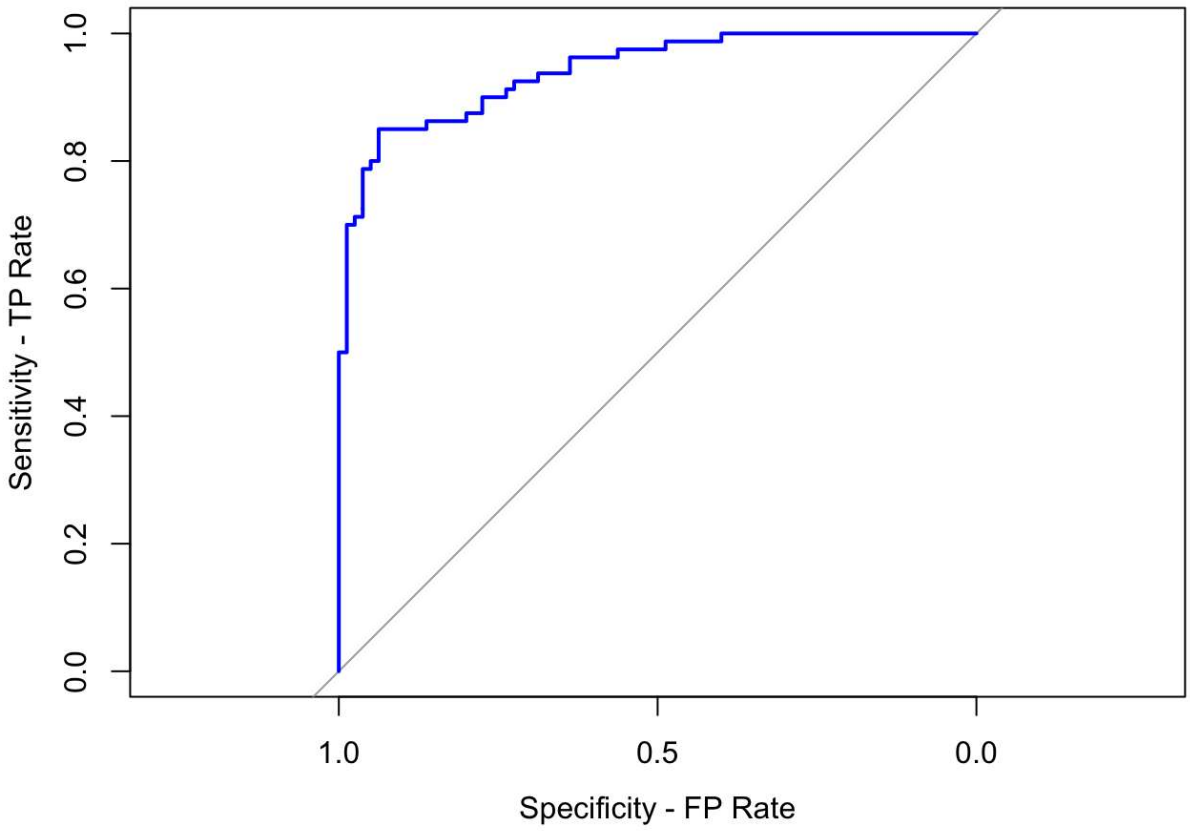

Figure 9: Receiver Operator Characteristics curve

The above plot depicts the trade-off between Sensitivity and Specificity. This classifier gives a curve closer to the top-left corner indicating a better performance. Usefulness based on AUC (Area Under the Curve) is as follows:

- 0.9 - 1 : excellent
- 0.8 - 0.9 : good
- 0.7 - 0.8 : fair
- 0.6 - 0.7 : poor
- 0.5 - 0.6 : failed

The area under the curve can be given by,

```
auc <- auc(ROC1)
knitr::kable(auc,col.names = "AUC")
```

AUC
0.9417188

AUC measures the entire two-dimensional area underneath the entire ROC curve. The area under our ROC curve is `0.9417188`.

Goal 2: What are the factors influencing the returns of betting companies?

Multiple Linear Regression to find variables associated with Betting rates.

From the preliminary analysis, we found that there is no correlation between the number of goals and returns of betting companies. Hence, this can be validated on further analysis. An assumption is that the returns from betting companies fluctuate for every match given the match statistics. We can find the factors that influence the returns of betting companies using multiple regression techniques.

Multiple linear regression is an extension of simple linear regression used to predict an outcome variable (y) based on the multiple distinct predictor variables (x) using the regression equation. The B365 betting returns of home team is taken into consideration for the analysis. It is compared against match statistics such as Full-time result, Half-time result, Home Team Goals, Goals, Shots, Shots on Target, Red cards and Yellow cards using multiple linear regression.

The variables containing match statistics and betting returns of other companies are stored in two seperate datasets like below.

```
#Filter all the numerical columns
laliga_num <- select_if(laliga_new,is.numeric)

# Data frame containing match statistics
laliga_game <- rbind(laliga_num[1:17])
names(laliga_game)
```

```
##  [1] "fult_home_goal"      "fult_away_goal"      "hlft_home_goal"
##  [4] "hlft_away_goal"      "home_shot"           "away_shot"
##  [7] "home_shot_on_target" "away_shot_on_target" "home_fouls"
## [10] "away_fouls"          "home_corners"        "away_corners"
## [13] "home_yellow"         "away_yellow"         "home_red"
## [16] "away_red"            "B365H"
```

```
# Data frame containing betting companies' information
laliga_comp <- rbind(laliga_num[17:34])
names(laliga_comp)
```

```
##  [1] "B365H" "B365D" "B365A" "BWH"   "BWD"   "BWA"   "IWH"   "IWD"   "IWA"
## [10] "PSH"   "PSD"   "PSA"   "WHH"   "WHD"   "WHA"   "VCH"   "VCD"   "VCA"
```

A multiple linear regression model for checking the influence on betting returns can be given by,

```
bet_fit <- lm(B365H ~ .,data=laliga_game)
summary(bet_fit)
```

```
##
## Call:
## lm(formula = B365H ~ ., data = laliga_game)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9994 -0.6263 -0.2473  0.3435  7.2053
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.1699416  0.8758093   4.761 4.22e-06 ***
## fult_home_goal      -0.1349159  0.1245902  -1.083  0.28046
## fult_away_goal       0.1708696  0.1599776   1.068  0.28706
## hlft_home_goal      -0.1194264  0.1674109  -0.713  0.47664
## hlft_away_goal       0.1347038  0.2118104   0.636  0.52569
## home_shot           -0.0498130  0.0329167  -1.513  0.13214
## away_shot            0.0009256  0.0315413   0.029  0.97662
## home_shot_on_target -0.0597099  0.0605478  -0.986  0.32552
## away_shot_on_target  0.1040065  0.0777437   1.338  0.18282
## home_fouls           0.0355979  0.0277389   1.283  0.20120
## away_fouls          -0.0836082  0.0291185  -2.871  0.00463 **
## home_corners        -0.0807554  0.0466661  -1.730  0.08543 .
## away_corners        -0.0191852  0.0529495  -0.362  0.71758
## home_yellow          0.0253346  0.0770141   0.329  0.74261
## away_yellow          0.0764062  0.0738125   1.035  0.30214
## home_red            -0.6415933  0.2881952  -2.226  0.02737 *
## away_red             0.2274102  0.3231432   0.704  0.48260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.337 on 163 degrees of freedom
## Multiple R-squared:  0.2663, Adjusted R-squared:  0.1943
## F-statistic: 3.698 on 16 and 163 DF,  p-value: 8.871e-06
```

From the above output, we can infer that

- The first step in interpreting the multiple regression analysis is to examine the F-statistic and the associated p-value, at the bottom of model summary.
- At least one of the match statistic variables(predictor) is significantly related to the betting return rates (outcome variable). (Given its p-value (8.87e-6) < Significance Level (0.05))
- The predictor variables shots on target (away),fouls (away) and home_red are significantly associated to betting returns of B365 company i.e., any changes made to those factors will affect the betting return rates.
- Other insignificant variables (without *) such as half time goals (home), goals, half time goals (away), home shots, away shots, fouls (home), etc. are not significantly associated to betting returns of B365 company (given its negative estimate value and Pr value) i .e., any changes made to those factors won't affect the betting returns.

- Around 19%(R-squared value) variation in betting returns can be explained by the significant factors found above, which is too less to be considered.
- The predictor variables that aren't significant can be removed and another model can be built.

The updated regression model can be given as,

```
bet_fit2 <- lm(B365H ~ away_fouls + home_red ,data=laliga_game)
summary(bet_fit2)
```

```
##
## Call:
## lm(formula = B365H ~ away_fouls + home_red, data = laliga_game)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6309 -0.8800 -0.3551  0.3188  7.3008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.12930    0.40403   7.745  7.1e-13 ***
## away_fouls  -0.03584    0.02797  -1.281    0.202
## home_red    -0.36822    0.30035  -1.226    0.222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.484 on 177 degrees of freedom
## Multiple R-squared:  0.01883,    Adjusted R-squared:  0.007739
## F-statistic: 1.698 on 2 and 177 DF,  p-value: 0.186
```

From the above output, we can infer that

- None of the predictor variables are significantly associated to the betting returns (Given the p-value,0.186 which is significantly greater than the significance level (0.05)).
- The R-squared values have drastically dropped to 0.7% i.e., nearly 0% indicating that none of the variables impact the betting return rates. Hence, this model can be omitted.

Our assumption on betting returns getting influenced by match statistics has been proved wrong. Let us validate if the returns of betting companies influence each other. A multiple linear regression model to compare the association between returns of B365 betting company for home team with other returns can be shown as,

```
fit3 <- lm(B365H ~ .,data=laliga_comp)
summary(fit3)
```

```
##
## Call:
## lm(formula = B365H ~ ., data = laliga_comp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45839 -0.03520  0.00133  0.04402  0.45410
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1854252  0.0617928  -3.001  0.00312 **
## B365D       -0.2042101  0.0742677  -2.750  0.00664 **
## B365A       -0.0024820  0.0226046  -0.110  0.91270
## BWH          0.2794360  0.0590445   4.733 4.80e-06 ***
## BWD          0.2767595  0.0928481   2.981  0.00332 **
## BWA          0.0397355  0.0375214   1.059  0.29117
## IWH         -0.1928660  0.1081414  -1.783  0.07638 .
## IWD         -0.0679355  0.0789997  -0.860  0.39109
## IWA         -0.0008805  0.0332919  -0.026  0.97893
## PSH         -0.0158331  0.1419639  -0.112  0.91134
## PSD          0.1055257  0.1384864   0.762  0.44717
## PSA          0.0099396  0.0569755   0.174  0.86173
## WHH          0.2472658  0.1343719   1.840  0.06757 .
## WHD          0.0941776  0.1321288   0.713  0.47701
## WHA         -0.0167770  0.0384479  -0.436  0.66316
## VCH          0.6984832  0.1074752   6.499 9.53e-10 ***
## VCD         -0.1491971  0.1444635  -1.033  0.30325
## VCA         -0.0416659  0.0335057  -1.244  0.21546
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1113 on 162 degrees of freedom
## Multiple R-squared:  0.995,  Adjusted R-squared:  0.9944
## F-statistic:  1878 on 17 and 162 DF,  p-value: < 2.2e-16
```

From the above output, we can infer that

- The predictor variables B365D (returns of Betting 365 company for draw), BWH (returns from B.W company for home) , BWD (returns from B.W company for draw) and VCH (returns from Viley company for home) are significantly associated to the betting returns of B365 company (home) thus denoting a strong relationship between them.
- The p-value is < 2.2e-16 and it is less than the significance level (0.05) denoting that at least one variable is significantly associated to the response variable. -The R-squared value is 0.9944 i.e., Almost 99% variation in betting returns can be explained by the significant variables found above.Thus, we can re-build another model omitting the insignificant variables in this model. The updated model can be given as,

```
fit4 <- lm(B365H ~ B365D + BWH + BWD + VCH,data=laliga_comp)
summary(fit4)
```

```
##
## Call:
## lm(formula = B365H ~ B365D + BWH + BWD + VCH, data = laliga_comp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46897 -0.03482  0.00290  0.04325  0.53043
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.14416    0.03553  -4.057 7.46e-05 ***
## B365D       -0.22014    0.05816  -3.785 0.000211 ***
## BWH          0.25927    0.04976   5.210 5.26e-07 ***
## BWD          0.23969    0.05858   4.092 6.52e-05 ***
## VCH          0.77550    0.05020  15.448  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.112 on 175 degrees of freedom
## Multiple R-squared:  0.9945, Adjusted R-squared:  0.9944
## F-statistic:  7880 on 4 and 175 DF,  p-value: < 2.2e-16
```

From the above output, we can infer that - All the predictor variables in the above model are significantly associated to the betting returns of B365 betting company (for home team). - Almost 99% variation in betting returns can be explained by all the variables in this model. - The regression equation can be given as B365H = - 0.22014*B365D* + *0.25927*BWH + 0.23969*BWD* + *0.77550* VCH - 0.14416. - A change in 2 units of betting returns can increase estimate value by two times if all the other values predictor variables are kept constant.

Thus, we can conclude that betting returns do not get impacted by match statistics. Instead, returns of Betting365 company (returns for home match) are influenced by returns of companies such as Betting 365 (returns for draw match), Betting World (returns for home match), Viley company (returns for home match).

Goal 3: Do all betting companies give the same returns?

Anova Test to compare the mean return rates of different betting companies

While placing bets it is confusing to choose a company that yields maximum results, hence differences in mean returns of companies are analysed. ANOVA test can be used to statistically validate if all companies provide similar returns.

```
laliga <- laliga_new

#State the hypothesis
#H0: Companies giving different returns
#H1: Companies giving same Returns

# Set significance level
alpha <- 0.05

# Dataframe for Betting companies
B365H <- data.frame('Returns' = c(laliga$B365H),
                    'companyname' = rep('B365H',180), stringsAsFactors = FALSE)


BWH <- data.frame('Returns' = c(laliga$BWH),
                      'companyname' = rep('BWH',180), stringsAsFactors = FALSE)


IWH <- data.frame('Returns' = c(laliga$IWH),
                      'companyname' = rep('IWH',180), stringsAsFactors = FALSE)


PSH <- data.frame('Returns' = c(laliga$PSH),
                      'companyname' = rep('PSH',180), stringsAsFactors = FALSE)

WHH <- data.frame('Returns' = c(laliga$WHH),
                      'companyname' = rep('WHH',180), stringsAsFactors = FALSE)


VCH <- data.frame('Returns' = c(laliga$VCH),
                      'companyname' = rep('VCH',180), stringsAsFactors = FALSE)


# Combine the Dataframe

Returnsdf <- c(B365H, BWH, IWH, PSH, WHH, VCH)


Returns <- rbind(B365H, BWH, IWH, PSH, WHH, VCH)
Returns$companyname <- as.factor(Returns$companyname)

# ANOVA Test
anova <- aov(Returns ~ companyname, data = Returns)

summary(anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## companyname    5    0.4  0.0749   0.036  0.999
## Residuals   1074 2255.1  2.0997
```

```
anovasum <- summary(anova)

#Find the critical value
qf(1 - 0.05,anovasum[[1]][1,1],anovasum[[1]][2,1])
```

```
## [1] 2.222434
```

```
criticalval <- 2.22

#Compute the test value.

test.value <- anovasum[[1]][[1,"F value"]]

test.value
```

```
## [1] 0.03565801
```

```
#Make the desicion
p.value <- anovasum[[1]][[1,"Pr(>F)"]]

ifelse(criticalval > 0.05,"Failed to reject Null Hypothesis","Reject Null Hypothesis")
```

```
## [1] "Failed to reject Null Hypothesis"
```

Observation: we have the resultant value which is greater than the significance level of 0.05, we fail to reject the null hypothesis because we do not have enough evidence to conclude that after statistically validating if all companies do not provide similar returns.

Conclusion

- Goal 1: The model to predict the Wins of the Home Team by predicting if the Full Time Results (fult_result) as H, was built by considering the variables (Half-Time Home Team Goals-hlft_home_goal, Away Team Shots on Target-away_shot_on_target and Red cards given to Home Team-home_red) that proved to be significant. This model showed an accuracy of 83.75% and has a confidence interval of (0.771, 0.891) at 95% confidence level. The true positive value is 69 whereas the true negative value is 65. The false positive value is 15 whereas the false negative is 11. In this case, false positive (Predicting ‘Away Team Win’ but actually ‘Home Team Win’) miscalculation is more damaging for the analysis.
- Goal 2: Using multiple regression, we were able to conclude that betting returns do not get impacted by match statistics. Instead it is influenced by betting returns of other companies i.e., Betting365 company (home team) was significantly associated to returns of companies such as B365 (draw), Betting World (home), Wiley-VC company (home).

- Goal 3: While validating if all companies provide similar returns using ANOVA test, we failed to reject the null hypothesis (p-value greater than the significance value) because we did not have enough evidence to conclude that the companies provide similar returns.

References

- (2022). Retrieved 31 January 2022, from https://cran.r-project.org/web/packages/ISLR/ISLR.pdf (https://cran.r-project.org/web/packages/ISLR/ISLR.pdf)
- Accuracy Vs Precision – NoSimpler. (2022). Retrieved 31 January 2022, from https://www.nosimpler.me/accuracy-precision/ (https://www.nosimpler.me/accuracy-precision/)
- set?, W. (2017). Why ROC Curve on test set?. Retrieved 31 January 2022, from https://stats.stackexchange.com/questions/265661/why-roc-curve-on-test-set (https://stats.stackexchange.com/questions/265661/why-roc-curve-on-test-set)
- What is a ROC Curve and How to Interpret It. (2018). Retrieved 31 January 2022, from https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/ (https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/)
- ROC curves – what are they and how are they used?. (2011). Retrieved 31 January 2022, from https://acutecaretesting.org/en/articles/roc-curves-what-are-they-and-how-are-they-used# (https://acutecaretesting.org/en/articles/roc-curves-what-are-they-and-how-are-they-used#):~:text=As%20 \ the%20area%20under%20an,stands%20for%20Receiver%20Operating%20Characteristic.