Capstone CYO - English Football Passing Analysis

Stephen Clarke

03/12/2020

Executive Summary

The purpose of this report was to create a model to evaluate the passing ability of football players in the English football leagues, via their pass accuracy, while minimising any biases which contribute to a higher or lower pass accuracy. Additionally, it provided a means of predicting a players pass accuracy, using the aforementioned biases.

Through data exploration and visualisation techniques, it was found that the pass accuracy of a player is impacted strongly by the position they play on the pitch and the team they play for. Using the average pass accuracy with the Naive Baseline Model, we could achieve a decent RMSE of **6.016**. The position bias was then accounted for and used to improve the RMSE of the models to **4.64**. This was then further improved by adding the team bias, to achieve an RMSE of **3.265**.

Regularisation was attempted with the number of 'Passes per 90' as a factor to try and further improve the model, but this unfortunately increased the RMSE to above 50, so would not be useful.

Method and Analysis

I have broken down our 'Method and Analysis' section into four sub-sections; Data Cleaning, Data Exploration, Insights Gained, and Modelling Approach.

Data Cleaning

To start, we load the required libraries, having previously installed them.

```
library(tidyverse)
library(dslabs)
library(data.table)
library(dplyr)
library(ggplot2)
library(ggthemes)
library(ggrepel)
ds_theme_set()
library(Lahman)
library(lattice)
library(e1071)
library(caret)
library(knitr)
library(tidyr)
library(stringr)
library(readr)
memory.limit(56000)
```

```
## [1] 56000
```

The data set which includes all of the passing data of football players from the top three divisions in English football is downloaded via GitHub. There are two blank columns in the Team_Passing_Data data set called 'X' and 'X.1', so these are removed along with the 'League' column, as this column already exists in the Player_Passing_Data data set.

```
Player_Passing_Data <- read.csv("https://raw.githubusercontent.com/sclarke-transferstat/HarvardX-Capstone-Passing-Analysis/main/English%20Football%20Player%20Passing%20Data%202019-20.csv")

Team_Passing_Data <- read.csv("https://raw.githubusercontent.com/sclarke-transferstat/HarvardX-Capstone-Passing-Analysis/main/English%20Football%20Team%20Passing%20Data%202019-20.csv")

#Remove League column in Team_Passing_Data as we already have that column in the Player_Passing Data. We also remove the last 2 columns as they are blank

Team_Passing_Data <- within(Team_Passing_Data, rm("League", "X", "X.1"))
```

We join these two data sets together by 'Team', using the left.join function, so that we can examine the players relative to the club that they have been playing for.

```
Passing_Data <- left_join(Player_Passing_Data, Team_Passing_Data, by = "Team")

kable(head(Passing_Data))
```

Player	Team	League	Position	Age	Market_value	Contract_expires	Matches_played	Minutes_played	Goals	xG	Birth_country	Passport_
A. Lacazette	Arsenal	Premier League	CF	29	4.2e+07	30/06/2022	30	2026	10	9.51	France	France, Guadeloup
B. Leno	Arsenal	Premier League	GK	28	3.2e+07	30/06/2023	30	2864	0	0.00	Germany	Germany
B. Saka	Arsenal	Premier League	LB, LWB, LAMF	19	4.0e+07	30/06/2024	26	1902	1	1.21	England	England, N
David Luiz	Arsenal	Premier League	LCB, CB	33	6.0e+06	30/06/2021	33	3038	2	1.66	Brazil	Brazil, Por
G. Xhaka	Arsenal	Premier League	LDMF, LCMF, DMF	28	2.8e+07	30/06/2023	31	2782	1	0.74	Switzerland	Switzerlan Albania
N. Pépé	Arsenal	Premier League	RAMF, RWF, CF	25	4.0e+07	30/06/2024	31	2172	5	4.23	Côte d'Ivoire	Côte d'Ivoi

```
Passing_Data$League <- factor(Passing_Data$League, levels=c("Premier League", "Championship", "League One"))
```

As the default ordering of character variables is alphabetical, we have to order the Leagues as factors based on their actual level. Therefore, we put the Premier League first, Championship second and League One third.

Data Exploration

We start by having a look at the outright most accurate and inaccurate passers in the Premier League.

```
Accurate_Passers <- Passing_Data %>%
  select(Player, Team, League, Position, Pass_Accuracy) %>%
  filter(League== "Premier League") %>%
  arrange(desc(Pass_Accuracy)) %>%
  head(n=15)

kable(Accurate_Passers)
```

Player	Team	League	Position	Pass_Accuracy
Rodri Hernández	Manchester City	Premier League	DMF, LDMF, RCMF	94.11
A. Christensen	Chelsea	Premier League	RCB, CB	93.02
I. Gündogan	Manchester City	Premier League	DMF, LCMF, LDMF	92.65
Kepa Arrizabalaga	Chelsea	Premier League	GK	92.45
C. Söyüncü	Leicester City	Premier League	LCB, LCB3	92.21
G. Wijnaldum	Liverpool	Premier League	LCMF3	92.02
K. Zouma	Chelsea	Premier League	LCB, RCB, CB	91.96
V. van Dijk	Liverpool	Premier League	LCB	91.77
Fernandinho	Manchester City	Premier League	LCB, RCB	91.70
Alisson	Liverpool	Premier League	GK	91.00
J. Gomez	Liverpool	Premier League	RCB	90.88
S. Papastathopoulos	Arsenal	Premier League	RCB, LCB	90.87
Ederson	Manchester City	Premier League	GK	90.61
B. Godfrey	Norwich City	Premier League	LCB, RCB	90.43
N. Aké	Bournemouth	Premier League	LCB, LCB3	90.41

It is clear that players playing as Centre Backs, Midfielders and Goalkeepers are more likely to have high passing accuracies. Additionally, most players in the top 15 tend to play for either Liverpool, Manchester City or Chelsea.

We explore the most inaccurate passers in the Premier League to see what could influence having a poor pass accuracy.

```
Inaccurate_Passers <- Passing_Data %>%
  select(Player, Team, League, Position, Pass_Accuracy) %>%
  filter(League== "Premier League") %>%
  arrange(Pass_Accuracy) %>%
  head(n=15)

kable(Inaccurate_Passers)
```

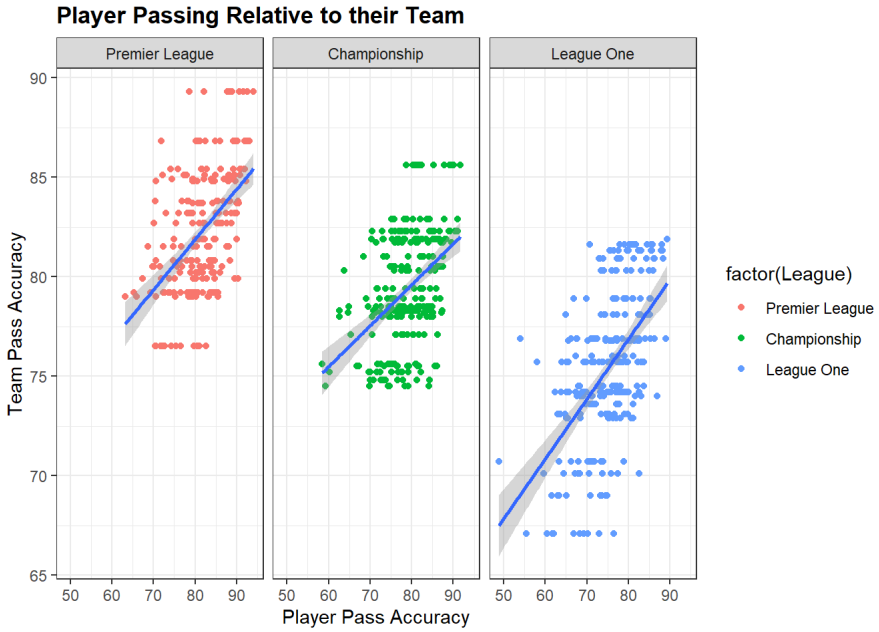
Player	Team	League	Position	Pass_Accuracy
S. Long	Southampton	Premier League	CF	63.21
T. Deeney	Watford	Premier League	CF	65.36
D. Ings	Southampton	Premier League	CF	65.93
C. Wilson	Bournemouth	Premier League	CF	67.36
Richarlison	Everton	Premier League	CF, RAMF, LW	68.67
A. Masina	Watford	Premier League	LB	69.35
F. Guilbert	Aston Villa	Premier League	RB, RWB	69.86
A. El Ghazi	Aston Villa	Premier League	RWF, RW, LW	70.04
Adama Traoré	Wolverhampton Wanderers	Premier League	RWF, RWB, CF	70.12
H. Kane	Tottenham Hotspur	Premier League	CF	70.41
C. Taylor	Burnley	Premier League	LB	70.44
P. Bardsley	Burnley	Premier League	RB	70.53
J. Vardy	Leicester City	Premier League	CF	70.54
Cédric Soares	Southampton	Premier League	RB, RWB	70.59
M. Antonio	West Ham United	Premier League	CF	70.67

We can see that Forwards and Full Backs tend to have the worst pass accuracy within the league. It is also noticeable that there are no Liverpool, Manchester City or Chelsea players within the bottom 15. This suggests that the position a player plays, and the team he plays for, have an impact on his pass accuracy. We will examine these potential ‘biases’ further.

Effect of Team

We use 'Team_Pass_Accuracy' to further analyse the effects of playing for a particular club, across the three leagues.

```
p <- Passing_Data %>%
group_by(League) %>%
ggplot(aes(Pass_Accuracy, Team_Pass_Accuracy)) +
geom_point(aes(col= factor(League))) +
geom_smooth(method = "lm")

p + facet_grid(cols = vars(League)) +
xlab("Player Pass Accuracy") +
  ylab("Team Pass Accuracy") +
  ggtitle("Player Passing Relative to their Team")
```



From the above graphs it certainly looks like the pass accuracy of a particular team, is correlated with the individual pass accuracy of players. We use the 'cor.test' function to get a value for this correlation and assess the statistical significance of this relationship.

```
cor.test(Passing_Data$Pass_Accuracy, Passing_Data$Team_Pass_Accuracy, method=c("pearson", "kendall", "spearman"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Passing_Data$Pass_Accuracy and Passing_Data$Team_Pass_Accuracy
## t = 20.505, df = 772, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5461909 0.6375941
## sample estimates:
##       cor
## 0.5938049
```

The very small p-value indicates statistical significance in this relationship.

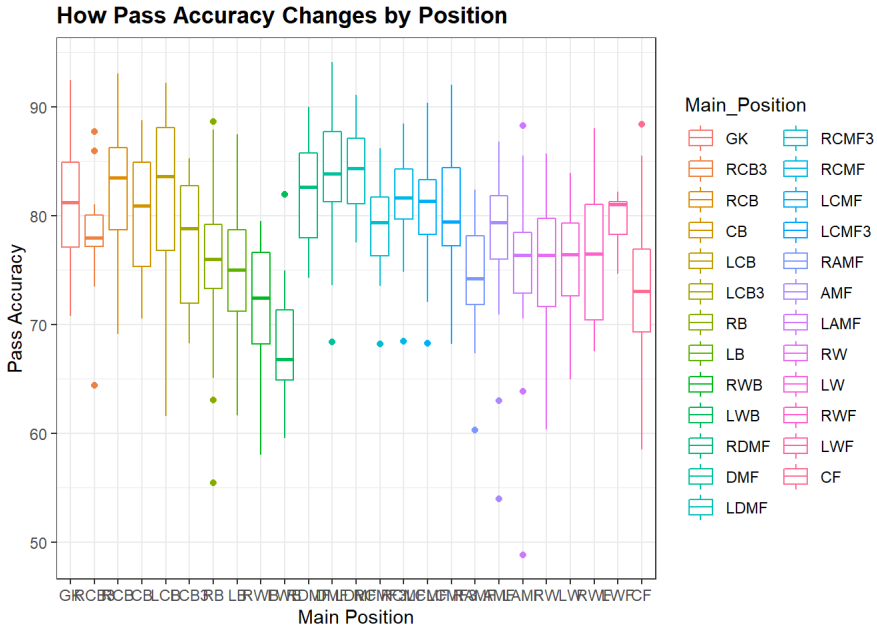
Effect of Position

We now look at the impact that playing in a certain position has on pass accuracy. To do so, we must first separate the 'Position' metric into "Main_Position", "Secondary_Position" and "Tertiary_Position", so that we can look at the position that a player has played in most during the season.

```
Passing_Data <- Passing_Data %>%
    separate(Position, c("Main_Position", "Secondary_Position", "Tertiary_Position"), fill = "right")

Passing_Data$Main_Position <- factor(Passing_Data$Main_Position, levels=c("GK", "RCB3", "RCB", "CB", "LCB", "LCB3", "RB", "L
B", "RWB", "LWB", "RDMF", "DMF", "LDMF", "RCMF3", "RCMF", "LCMF", "LCMF3", "RAMF", "AMF", "LAMF", "RW", "LW", "RWF", "LWF",
"CF"))
```

Once the data has been split into the 'Main Position' of the players, these positions must then be ordered. Otherwise, they will be in alphabetical order by default. This data can then be better visualised in a boxplot.

```
Passing_Data %>% ggplot(aes(x = Main_Position, y = Pass_Accuracy, col = Main_Position)) +
     geom_boxplot() +
xlab("Main Position") +
  ylab("Pass Accuracy") +
  ggtitle("How Pass Accuracy Changes by Position")
```

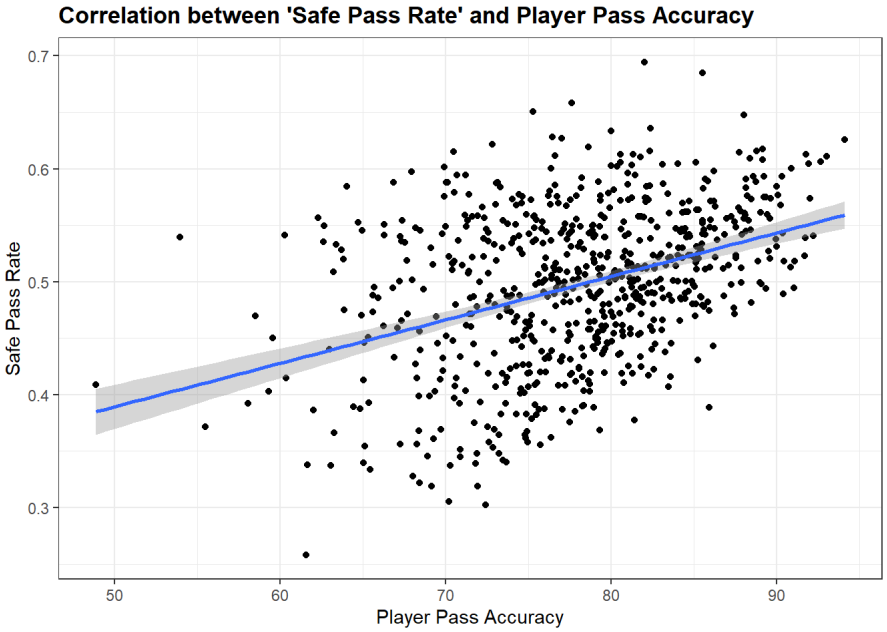


The boxplot above gives a very good indication of the varying pass accuracy of different positions. Centre Backs and Defensive Midfielders have the highest median pass accuracy, while Centre Forwards and Full Backs have the lowest.

Effect of Pass Difficulty

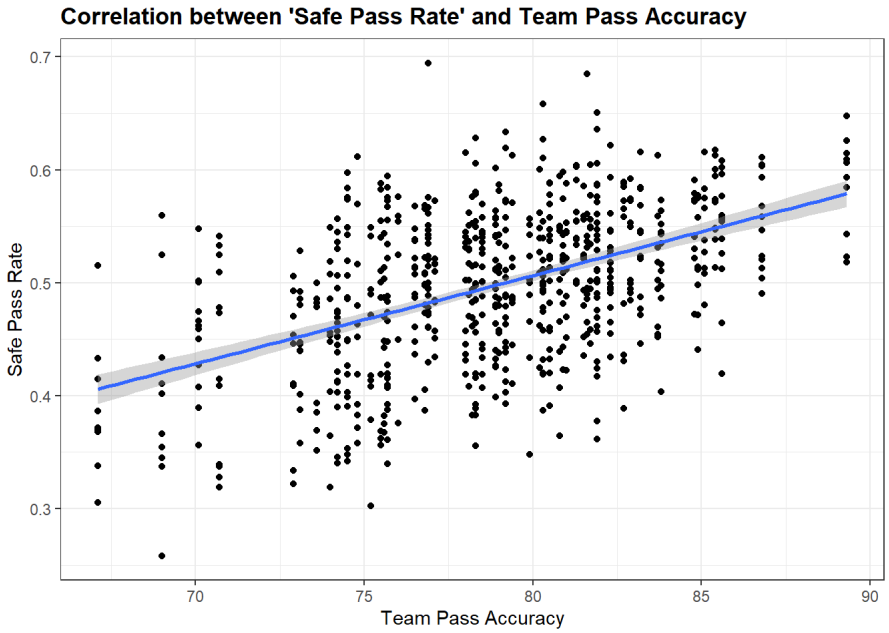
Another factor in pass accuracy I would like to consider is the difficulty of passes. I have created a new column called 'Safe_Pass_Rate' which is how many backward and lateral passes a player attempts, relative to their total number of passes attempted. As goalkeepers play a very small number of backward and lateral passes, I have removed them from the list.

```
Passing_Data <- Passing_Data %>%
  mutate(Safe_Pass_Rate = (Back_passes_per_90 + Lateral_passes_per_90)/Passes_per_90) %>%
  filter(Main_Position %in% c("RCB3", "RCB", "CB", "LCB", "LCB3", "RB", "LB", "RWB", "LWB", "RDMF", "DMF", "LDMF", "RCMF3",
"RCMF", "LCMF", "LCMF3", "RAMF", "AMF", "LAMF", "RW", "LW", "RWF", "LWF", "CF"))

Passing_Data %>%
  ggplot(aes(Pass_Accuracy, Safe_Pass_Rate)) +
  geom_point() +
  geom_smooth(method = "lm") +
xlab("Player Pass Accuracy") +
  ylab("Safe Pass Rate") +
  ggtitle("Correlation between 'Safe Pass Rate' and Player Pass Accuracy")
```



It seems that there is certainly a correlation between how many 'Safe Passes' a player makes and his pass accuracy, but this could be explained by the team that they are playing for. Teams like Liverpool and Man City play a large number of lateral passes due to their high possession numbers, so this could be what really provides the correlation.

```
Passing_Data %>%
  ggplot(aes(Team_Pass_Accuracy, Safe_Pass_Rate)) +
  geom_point() +
  geom_smooth(method = "lm") +
xlab("Team Pass Accuracy") +
  ylab("Safe Pass Rate") +
  ggtitle("Correlation between 'Safe Pass Rate' and Team Pass Accuracy")
```



The above graph between Team_Pass_Accuracy and Safe_Pass_Rate shows a similar correlation with seemingly smaller error. Furthermore, making a large number of safe passes, does not mean that a player isn't good at making 'non-safe' passes.

Insights Gained

Through the Data Exploration section, we have looked at numerous factors that seem to influence the passing accuracy of an individual footballer. The position that a player plays seems to have a large effect on their pass accuracy, as does the team that they are playing for. The 'main' position and team will be the first things we look at when trying to get a predictive RMSE for our players. For example, a central midfielder for Liverpool, such as Georginio Wijnaldum, will have a much higher pass accuracy than a right back for Burnley, such as Phil Bardsley. Therefore, we will account for position bias and team bias in our Modelling Approach section.

Modelling Approach

The seed is set to 1 and the data is split into test and train sets, where the train set is 90% of Passing_Data.

```
set.seed(1, sample.kind="Rounding")

train_index <- createDataPartition(y = Passing_Data$Pass_Accuracy, times = 1, p = 0.9, list = FALSE)

Passing_Data_train <- Passing_Data[train_index,]
Passing_Data_test <- Passing_Data[-train_index,]
```

Below is the standard formula for calculating root mean squared error (RMSE), which assesses the effectiveness of a predictive model.

```
RMSE <- function(true_ratings, predicted_ratings){
  sqrt(mean((true_ratings - predicted_ratings)^2, na.rm = TRUE))
}
```

We would like to get as low an RMSE as possible for the model to be truly predictive.

Naive Baseline RMSE

We start our approach to working out the RMSE using the Naive Baseline Model.

The formula for this is:

$$Y_{u,i} = \hat{\mu} + \varepsilon_{u,i}$$

Where $\hat{\mu}$ is the mean and $\varepsilon_{i,u}$ is the independent errors centered at 0.

The average pass accuracy for all players in the data set is as below:

```
mu <- mean(Passing_Data_train$Pass_Accuracy)
mu
```

```
## [1] 77.97713
```

```
naive_rmse <- RMSE(Passing_Data_test$Pass_Accuracy, mu)

naive_rmse
```

```
## [1] 6.016368
```

```
model_results <- data.frame(model = "Naive Baseline Model", RMSE = naive_rmse)

model_results
```

```
##                 model     RMSE
## 1 Naive Baseline Model 6.016368
```

The RMSE of the Naive Baseline Model is **6.016**, which is not a bad RMSE value, but this could be improved upon.

Using Position Bias to Improve RMSE

Earlier the data was partioned into test and train sets to allow for more in depth modelling. These are called Passing_Data_test and Passing_Data_train respectively.

We add the term b_P to compensate for position bias, where b_P is the average pass accuracy for each given position in this data set. The new formula will be:

$$Y_{u,i} = \mu + b_P + \varepsilon_{u,i}$$

To get b_P we use the least squares estimate.

```
Position_avg <- Passing_Data_train %>%
   group_by(Main_Position) %>%
   summarise(b_P = mean(Pass_Accuracy - mu))
```

```
Position_avg %>% arrange(desc(b_P))
```

```
## # A tibble: 24 x 2
##    Main_Position   b_P
##    <fct>         <dbl>
##  1 LDMF           5.90
##  2 DMF            5.81
##  3 RDMF           5.18
##  4 RCB            4.31
##  5 LCB            4.12
##  6 RCMF           3.59
##  7 LCMF           2.86
##  8 CB             2.13
##  9 LCMF3          1.97
## 10 LWF            1.57
## # ... with 14 more rows
```

We see from the above table how some positions have much higher pass accuracies than others. This is what we are trying to account for.

We then use our RMSE formula to see how effective it is to account for position bias, and hence, how good this method is for prediction.

```
predicted_ratings <- Passing_Data_test %>%
  left_join(Position_avg, by = 'Main_Position') %>%
  mutate(pred1 = mu + b_P) %>%
  .$pred1

#.$ means pull() function

Position_bias <- RMSE(Passing_Data_test$Pass_Accuracy, predicted_ratings)
```

```
model_results <- bind_rows(model_results, data_frame(model = "Position Bias Model", RMSE = Position_bias))

model_results
```

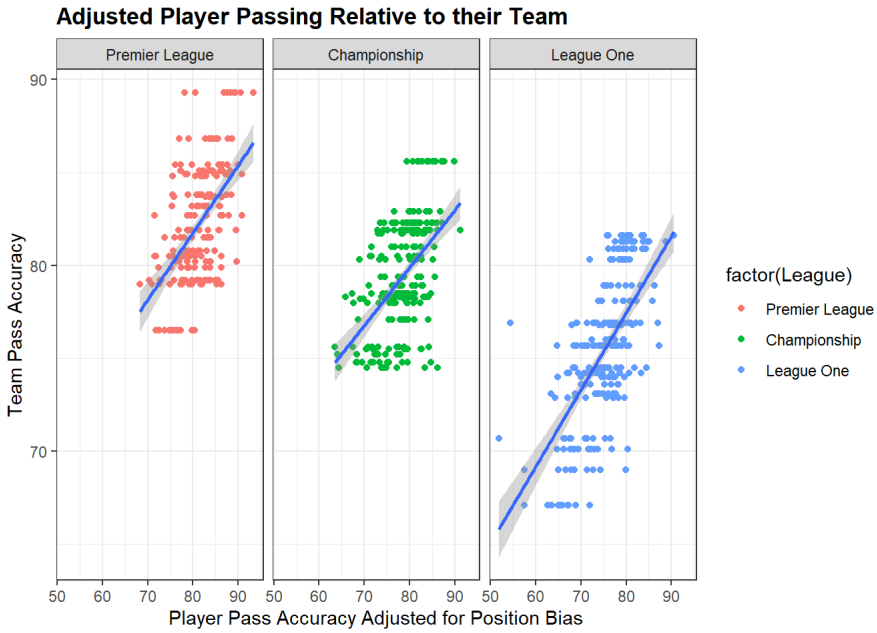
```
##                 model     RMSE
## 1 Naive Baseline Model 6.016368
## 2  Position Bias Model 4.639758
```

The RMSE has been reduced to just under **4.64** by including position bias.

We join our original Passing_Data with the Position_avg, and we can then visualise the correlation between Team Pass Accuracy and Player Pass Accuracy, once the position bias has been accounted for.

```
Passing_Data_With_PBias <- left_join(Passing_Data, Position_avg, by = "Main_Position")

Passing_Data_With_PBias <- Passing_Data_With_PBias %>% mutate(New_Pass_Acc = Pass_Accuracy - b_P) %>% select(Player, Main_Po
sition, Passport_country, Team, League, Minutes_played, Pass_Accuracy, Team_Pass_Accuracy, b_P, New_Pass_Acc) %>% arrange(de
sc(New_Pass_Acc))

p <- Passing_Data_With_PBias %>%
group_by(League) %>%
ggplot(aes(New_Pass_Acc, Team_Pass_Accuracy)) +
geom_point(aes(col= factor(League))) +
geom_smooth(method = "lm")

p + facet_grid(cols = vars(League)) +
xlab("Player Pass Accuracy Adjusted for Position Bias") +
  ylab("Team Pass Accuracy") +
  ggtitle("Adjusted Player Passing Relative to their Team")
```

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



The correlation looks to be stronger now that position bias has been taken into account and there appears to be a clear indication that Team bias should also be accounted for.

Using Team Bias to Improve RMSE

The next factor to consider to improve the RMSE, is the team that a player plays for. Therefore, we must add in another term, b_T, to compensate for this.

$$Y_{u,i} = \mu + b_P + b_T + \varepsilon_{u,i}$$

Where b_T is the team bias.

```
Team_avg <- Passing_Data_train %>%
  left_join(Position_avg, by = 'Main_Position') %>%
  group_by(Team) %>%
  summarize(b_T = mean(Pass_Accuracy - mu - b_P))
```

```
Team_avg
```

```
## # A tibble: 68 x 2
##    Team                  b_T
##    <fct>               <dbl>
##  1 Accrington Stanley -3.93
##  2 AFC Wimbledon      -6.91
##  3 Arsenal             5.96
##  4 Aston Villa        -0.106
##  5 Barnsley           -2.13
##  6 Birmingham City    -3.78
##  7 Blackburn Rovers    0.982
##  8 Blackpool          -2.25
##  9 Bolton Wanderers   -1.22
## 10 Bournemouth         1.28
## # ... with 58 more rows
```

From the tibble above, we can see the effect of team bias. An Arsenal player will have on average a 5.96 higher passing accuracy than the average player, while an AFC Wimbledon will have 6.91 lower passing accuracy.

We then use our RMSE formula to see how effective this method is for prediction.

```
predicted_ratings <- Passing_Data_test %>%
  left_join(Position_avg, by = 'Main_Position') %>%
  left_join(Team_avg, by = 'Team') %>%
  mutate(b_P_b_T = mu + b_P + b_T) %>%
  .$b_P_b_T

Team_And_Position_bias <- RMSE(Passing_Data_test$Pass_Accuracy, predicted_ratings)
```

```
model_results <- bind_rows(model_results, data_frame(model = "Position and Team Bias Model", RMSE = Team_And_Position_bias))

model_results
```

```
##                          model     RMSE
## 1         Naive Baseline Model 6.016368
## 2          Position Bias Model 4.639758
## 3 Position and Team Bias Model 3.265389
```

By including a factor for the bias of teams, we have reduced the RMSE down to **3.265**. This is quite a good RMSE that would suggest that this method would predict a player's passing accuracy relatively accurately.

We join our original Passing_Data with the Position_avg and Team_avg to see who would be considered the best passers in the Premier League if we accounted entirely for position and team, and we can then visualise the correlation between Team Pass Accuracy and Player Pass Accuracy, once the position bias has been accounted for.

```
Passing_Data_With_PBias <- left_join(Passing_Data, Position_avg, by = "Main_Position")

Passing_Data_With_PTBias <- left_join(Passing_Data_With_PBias, Team_avg, by = "Team")

Passing_Data_With_PTBias <- Passing_Data_With_PTBias %>%
  mutate(New_Pass_Acc = Pass_Accuracy - b_P - b_T) %>%
  select(Player, Main_Position, Passport_country, Team, League, Minutes_played, Pass_Accuracy, Passes_per_90, Team_Pass_Accu
racy, b_P, b_T, New_Pass_Acc) %>% arrange(desc(New_Pass_Acc))

PL_Best_Passers <- Passing_Data_With_PTBias %>%
  filter(League == "Premier League") %>%
select(Player, Main_Position, Team, Pass_Accuracy, Passes_per_90, b_P, b_T, New_Pass_Acc) %>%
 head(n=20)

kable(PL_Best_Passers)
```

Player	Main_Position	Team	Pass_Accuracy	Passes_per_90	b_P	b_T	New_Pass_Acc
J. Grealish	LW	Aston Villa	83.91	33.49	-2.319303	-0.1061795	86.33548
Joelinton	CF	Newcastle United	84.64	21.03	-5.037579	3.6234314	86.05415
Jonny Otto	LWB	Wolverhampton Wanderers	81.93	31.76	-8.987129	4.9863675	85.93076
W. Smallbone	RW	Southampton	83.17	43.64	-3.131929	0.4831613	85.81877
L. Shaw	LB	Manchester United	87.47	54.14	-3.367314	5.1667513	85.67056
N. Aké	LCB	Bournemouth	90.41	37.69	4.120239	1.2754186	85.01434
J. Ayew	CF	Crystal Palace	84.50	20.19	-5.037579	4.6830972	84.85448
M. Obafemi	CF	Southampton	80.15	16.34	-5.037579	0.4831613	84.70442
Heung-Min Son	LAMF	Tottenham Hotspur	85.53	26.85	-2.986360	4.1847261	84.33163
Gabriel Jesus	CF	Manchester City	88.41	23.68	-5.037579	9.1318720	84.31571
G. Wijnaldum	LCMF3	Liverpool	92.02	45.25	1.967315	5.9581054	84.09458
C. Söyüncü	LCB	Leicester City	92.21	59.42	4.120239	4.3335718	83.75619
Gerard Deulofeu	CF	Watford	78.63	29.37	-5.037579	0.1841939	83.48338
H. Winks	LCMF	Tottenham Hotspur	90.37	59.79	2.859121	4.1847261	83.32615
L. Dendoncker	RCB3	Wolverhampton Wanderers	87.69	36.11	-0.271296	4.9863675	82.97493
A. Christensen	RCB	Chelsea	93.02	57.97	4.307277	5.8680445	82.84468
D. Rice	RCMF	West Ham United	88.44	40.04	3.588032	2.3582243	82.49374
A. Smith	RB	Bournemouth	81.76	33.70	-1.986038	1.2754186	82.47062
S. Kolašinac	LB	Arsenal	84.91	43.69	-3.367314	5.9635859	82.31373
J. Rodriguez	CF	Burnley	75.36	17.24	-5.037579	-1.6607716	82.05835

We see that there is now a much wider variety in position and team for the ‘Top 20 Passers’ in the Premier League.

This can be specified further to get more insights for more niche players. For example, if we were looking for the best Irish passers who are currently playing regularly in English football's top two tiers.

```
Irish_Best_Passers <- Passing_Data_With_PTBias %>% filter(str_detect(Passport_country, "Republic of Ireland"), League %in% c
("Premier League", "Championship"), Minutes_played >= 1000) %>%
  select(Player, Main_Position, Passport_country, Team, League, Pass_Accuracy, Passes_per_90, b_P, b_T, New_Pass_Acc) %>%
  head(n=20)

kable(Irish_Best_Passers)
```

Player	Main_Position	Passport_country	Team	League	Pass_Accuracy	Passes_per_90	b_P	b_T	New_Pass_Ac
D. Crowley	RW	England, Republic of Ireland	Birmingham City	Championship	81.15	39.36	-3.1319293	-3.7804643	88.0623
J. Grealish	LW	England, Republic of Ireland	Aston Villa	Premier League	83.91	33.49	-2.3193033	-0.1061795	86.3354
J. Cullen	RCMF3	England, Republic of Ireland	Charlton Athletic	Championship	84.58	38.30	0.4671564	0.2439461	83.8689
J. Molumby	RCMF	Republic of Ireland	Millwall	Championship	86.03	36.30	3.5880320	-1.3933702	83.8353
R. Stearman	RCB	England, Republic of Ireland	Huddersfield Town	Championship	86.86	33.98	4.3072774	-0.2787062	82.8314
D. Rice	RCMF	England, Republic of Ireland	West Ham United	Premier League	88.44	40.04	3.5880320	2.3582243	82.4937
D. McGoldrick	CF	England, Republic of Ireland	Sheffield United	Premier League	80.57	26.30	-5.0375788	3.5675715	82.0400
D. Nugent	CF	England, Republic of Ireland	Preston North End	Championship	77.99	21.85	-5.0375788	1.2866133	81.7409
S. Hogan	CF	England, Republic of Ireland	Birmingham City	Championship	72.60	10.14	-5.0375788	-3.7804643	81.4180
J. Collins	CF	England, Republic of Ireland	Luton Town	Championship	74.45	18.41	-5.0375788	-1.7068609	81.1944
M. Doherty	RWB	Republic of Ireland	Wolverhampton Wanderers	Premier League	79.48	40.55	-6.6383793	4.9863675	81.1320
M. Keane	RCB	England, Republic of Ireland	Everton	Premier League	87.54	42.53	4.3072774	2.1898807	81.0428
K. Naughton	RB	England, Republic of Ireland	Swansea City	Championship	81.42	44.33	-1.9860384	2.4552026	80.9508
A. Pearce	CB	Scotland, Republic of Ireland	Millwall	Championship	81.38	26.37	2.1255979	-1.3933702	80.6477
E. Stevens	LWB	Republic of Ireland	Sheffield United	Premier League	74.97	39.02	-8.9871293	3.5675715	80.3895
S. Coleman	RB	Republic of Ireland	Everton	Premier League	80.46	37.59	-1.9860384	2.1898807	80.2561
S. Williams	LCMF	Republic of Ireland	Millwall	Championship	81.72	32.76	2.8591207	-1.3933702	80.2542
C. O'Dowda	LW	England, Republic of Ireland	Bristol City	Championship	76.97	27.69	-2.3193033	-0.6705319	79.9598
P. Bamford	CF	England, Republic of Ireland	Leeds United	Championship	75.89	13.71	-5.0375788	1.0638327	79.8637
C. Clark	LCB3	England, Republic of Ireland	Newcastle United	Premier League	81.82	21.74	-1.1471293	3.6234314	79.3437

Regularisation

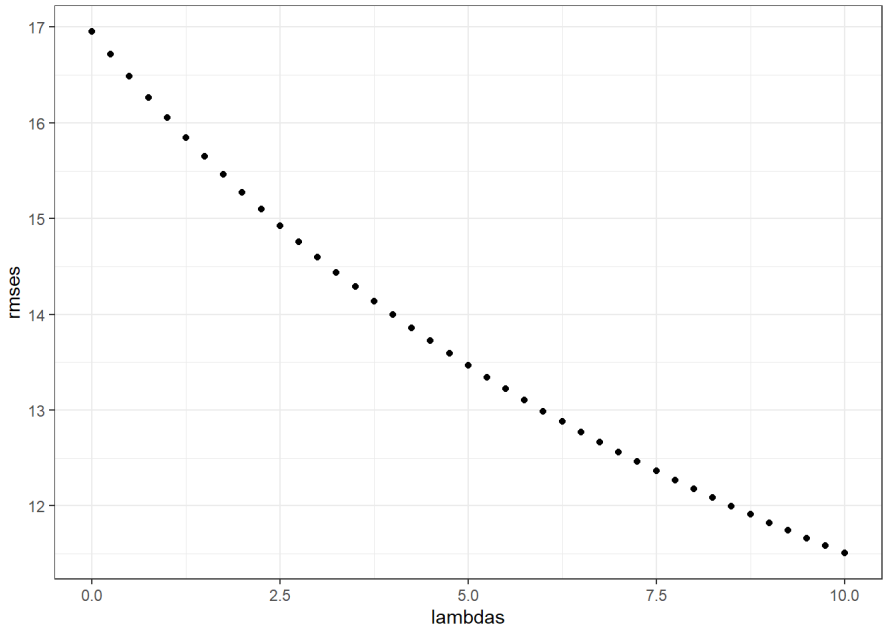
When we put the Players in order of the newly accounted for pass accuracy, we notice that some of the players who have benefitted most from this model, have a relatively low 'Passes_per_90'. Therefore, it is worth looking into if we could improve our RMSE by accounting for this. To do so, we will use Regularisation.

Regularisation allows us to penalise players that have got high 'New_Pass_Acc' but low 'Passes_per_90'.

We use λ as a 'tuning parameter' and use cross-validation to get its value.

```
lambdas <- seq(0, 10, 0.25)

mu <- mean(Passing_Data_train$Pass_Accuracy)
just_the_sum <- Passing_Data_train %>%
  group_by(Main_Position) %>%
  summarize(s = sum(Pass_Accuracy - mu), n_P = Passes_per_90)


rmses <- sapply(lambdas, function(l){
  predicted_ratings <- Passing_Data_test %>%
    left_join(just_the_sum, by='Main_Position') %>%
    mutate(b_P = s/(n_P+l)) %>%
    mutate(pred = mu + b_P) %>%
    pull(pred)
  return(RMSE(predicted_ratings, Passing_Data_test$Pass_Accuracy))
})
```

```
qplot(lambdas, rmses)
```




```
lambdas[which.min(rmses)]
```

```
## [1] 10
```

```
min(rmses)
```

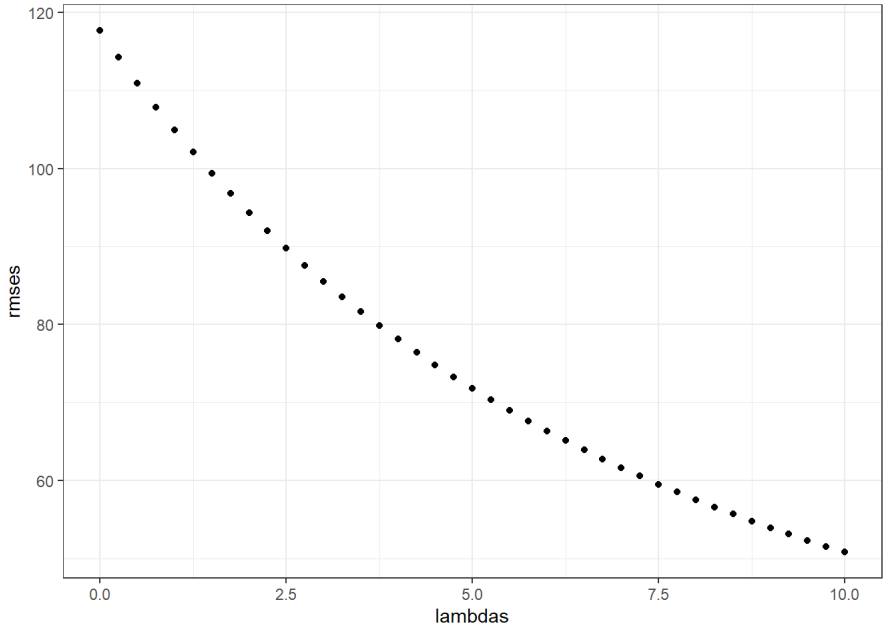
```
## [1] 11.50494
```

The above graph shows the lambda value at which the RMSE is at its lowest, having regularised for position bias. However, the RMSE here looks higher than our achieved RMSE without regularisation. This could be a result of overtraining.

We want to account for both position bias and team bias, so we will see if this improves or gets worse through further analysis.

```
lambdas <- seq(0, 10, 0.25)

rmses <- sapply(lambdas, function(l){

  mu <- mean(Passing_Data_train$Pass_Accuracy)

  b_P <- Passing_Data_train %>%
    group_by(Main_Position) %>%
    summarize(b_P = sum(Pass_Accuracy - mu)/(Passes_per_90 +l))

  b_T <- Passing_Data_train %>%
    left_join(b_P, by="Main_Position") %>%
    group_by(Team) %>%
    summarize(b_T = sum(Pass_Accuracy - b_P - mu)/(Passes_per_90 +l))

  predicted_ratings <-
    Passing_Data_test %>%
    left_join(b_P, by = "Main_Position") %>%
    left_join(b_T, by = "Team") %>%
    mutate(pred = mu + b_P + b_T) %>%
    pull(pred)

return(RMSE(predicted_ratings, Passing_Data_test$Pass_Accuracy))
})
```

```
qplot(lambdas, rmses)
```




```
lambda <- lambdas[which.min(rmses)]
lambda
```

```
## [1] 10
```

```
model_results <- bind_rows(model_results, data_frame(model = "Regularised Position and Team Bias Model", RMSE = min(rmses)))

model_results
```

```
##                                    model      RMSE
## 1                    Naive Baseline Model  6.016368
## 2                     Position Bias Model  4.639758
## 3             Position and Team Bias Model  3.265389
## 4 Regularised Position and Team Bias Model 50.799410
```

The RMSE rises further to **50.799** so it would seem that the regularisation technique used has overtrained the model.

Results

The below model results show us that the best way to get a small RMSE value from the methods attempted, is by using the bias of the position and team in which a player plays.

```
model_results
```

```
##                                     model      RMSE
## 1                     Naive Baseline Model  6.016368
## 2                      Position Bias Model  4.639758
## 3             Position and Team Bias Model  3.265389
## 4 Regularised Position and Team Bias Model 50.799410
```

We have achieved an RMSE of **3.265** with the 'Position and Team Bias Model' which makes it predictive.

Conclusion

I was able to provide an RMSE of **3.265**, which would mean that this system would be provide a solid prediction of a player's pass accuracy based on his team and position. This could be used to evaluate the passing abilities of players more fairly, who may play in teams or positions that are not condusive to a high passing accuracy. What must be taken into consideration, however, is the question; is a player's passing great because he plays for a better team, or does he play for a better team because he is a great passer? Should Liverpool and Man City players be penalised for playing in better teams, as many may be in those teams due to this excellent pass accuracy, and will contribute to the team having a high overall passing accuracy. To improve the RMSE number further, it would be good to take into account the difficulty of players' passes, although this is something difficult to quantify using only some of the simplistic metrics that we have to work with. In future analyses, I would like to use a larger data set with numerous leagues worldwide, providing a data set of thousands of football players to work with, and get more reliable answers.