AIT 580 Data Analysis Individual Project
By Kamil Ismailov

Introduction

The dataset which I chose contains information about the results of Premier League season 2018-2019 [1]. All abbreviations of the column are presented in Appendix A. The English soccer tournament contains twenty teams that play each other two times: at home and away. So, the total number of games is 380 per season. According to the official table results, the Manchester City soccer club became a champion in the season 2018-2019 because they scored the most points [2]. The main rival for the Manchester City was Liverpool in this season. The purpose of this project has two goals. The first aim is to use the selected dataset and compare the performance of these two soccer clubs. The second goal is to build a prediction method by using the results of matches and check the model by comparing the last game week. The unit of analysis is a number of scored and conceded goals because it is the most important element of the game.

Summary statistics of the match results

We need to make summary statistics of goals for each team to understand better the clubs' performance. The best and easiest way to do is to use pgAdmin 4 software. This PostgreSQL tool helps to work with queries. First, I load the dataset in the pgAdmin 4 Database (figure 1) and extract the table with information about the home and away team performance. Then, I exclude the date column because it had an inappropriate format for this software.

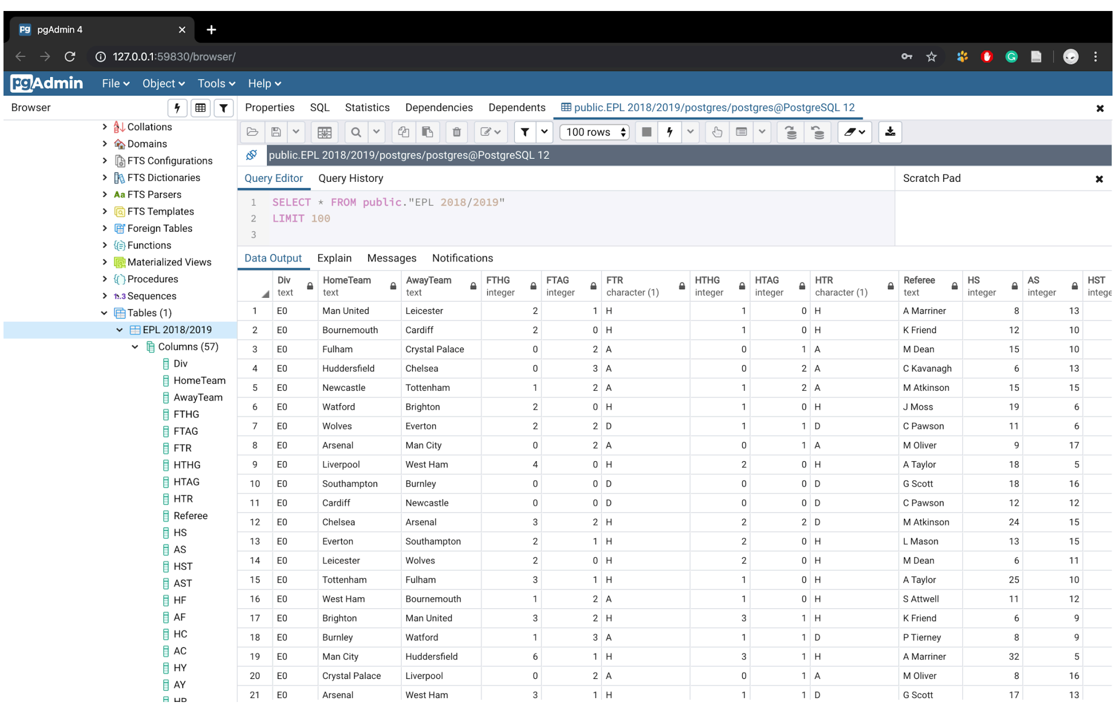


Figure 1. EPL 2018/2019 table in pgAdmin 4

HomeTeam	avg_goals	avg_half_time_goals	avg_conceded_goals	avg_HT_conceded_goals	sum_goals	avg_shots	avg_shots_on_target
Man City	3.00	1.47	0.63	0.37	57	20.32	7.79
Liverpool	2.89	1.37	0.53	0.26	55	17.63	6.63
Arsenal	2.21	0.74	0.84	0.42	42	13.47	5.05
Chelsea	2.05	0.84	0.63	0.32	39	17.11	6.21
Tottenham	1.79	0.68	0.84	0.26	34	16.21	5.53
Man United	1.74	0.84	1.32	0.37	33	14.95	6.63
West Ham	1.68	0.63	1.42	0.68	32	12.89	4.63
Everton	1.58	0.74	1.11	0.58	30	14.68	4.47
Bournemouth	1.58	0.79	1.32	0.53	30	12.05	4.32
Wolves	1.47	0.53	1.11	0.47	28	14.05	4.47
Southampton	1.42	0.74	1.58	0.68	27	13.58	4.68
Watford	1.37	0.37	1.47	0.74	26	11.79	4.37
Burnley	1.26	0.74	1.68	0.63	24	11.05	3.47
Leicester	1.26	0.53	1.05	0.63	24	15.68	5.21
Newcastle	1.26	0.53	1.32	0.63	24	14.00	4.21
Fulham	1.16	0.47	1.89	1.05	22	14.42	4.53
Cardiff	1.11	0.42	2.00	0.95	21	12.11	3.58
Brighton	1.00	0.53	1.47	0.63	19	10.53	2.79
Crystal Palace	1.00	0.42	1.21	0.32	19	15.47	4.00
Huddersfield	0.53	0.21	1.63	0.95	10	10.68	3.00

Table 1. Home Team Performance

According to Table 1, Manchester City superior in all respects at the home stadium. Their average numbers of shots, shots of a target, and scored goals in both halt time and full time are the highest. However, the average number of conceded goals of Liverpool FC is the lowest, but Manchester City scored more.

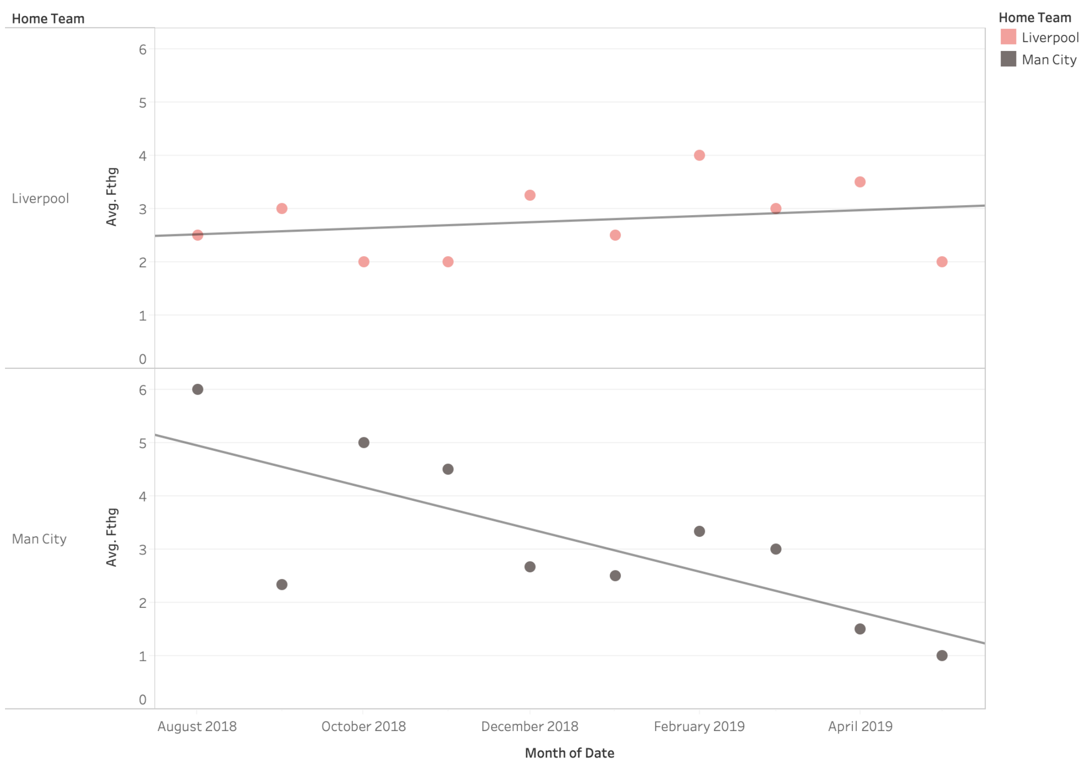
AwayTeam	avg_goals	avg_HT_goals	avg_conceded_goals	avg_HT_conceded_goals	sum_goals	avg_shots	avg_shots_on_target
Man City	2.00	1.11	0.58	0.21	38	15.63	5.89
Liverpool	1.79	0.63	0.63	0.26	34	12.58	5.26
Tottenham	1.74	0.95	1.21	0.32	33	12.00	4.42
Crystal Palace	1.68	0.63	1.58	0.37	32	10.37	3.68
Man United	1.68	1.00	1.53	0.84	32	12.74	5.21
Arsenal	1.63	0.74	1.84	1.05	31	11.05	3.89
Leicester	1.42	0.26	1.47	0.79	27	11.42	4.47
Bournemouth	1.37	0.68	2.37	1.21	26	11.47	4.26
Watford	1.37	0.58	1.63	0.63	26	11.05	3.58
Chelsea	1.26	0.58	1.42	0.58	24	14.79	4.21
Everton	1.26	0.58	1.32	0.32	24	11.37	4.42
Burnley	1.11	0.53	1.89	0.89	21	7.84	2.58
West Ham	1.05	0.53	1.47	0.58	20	10.32	3.63
Wolves	1.00	0.37	1.32	0.68	19	10.89	3.47
Southampton	0.95	0.42	1.84	0.89	18	11.68	3.84
Newcastle	0.95	0.68	1.21	0.42	18	9.37	3.26
Brighton	0.84	0.37	1.68	0.74	16	8.68	2.89
Cardiff	0.68	0.16	1.63	0.63	13	9.84	3.05
Huddersfield	0.63	0.37	2.37	1.05	12	10.37	3.21
Fulham	0.63	0.32	2.37	1.11	12	9.42	3.32

Table 2. Away Team Performance

According to Table 2, there is also domination by Manchester City in away games. Now, it becomes clear why Manchester City is a champion of season 2018-2019. They scored more, conceded less, and shot on target more frequently.

Visualization of the match results

As we want to compare the result of Manchester City and Liverpool FC, the represented table is not a good way because their numbers are very close to each other. In this case, a visualization of the tables can show the difference between the clubs' performance better. Tableau is the convenient software for data visualization. We use a scatterplot to show average goals scored and conceded per month and look through the tendency.



The plot of average of Fthg for Date Month broken down by Home Team. Color shows details about Home Team. The view is filtered on Home Team, which keeps Liverpool and Man City.

Figure 2. The average number of goals scored at Home (Liverpool FC vs. Man City)

Figure 2 shows that Manchester City started the season at home stadium much better than Liverpool FC. The average number of goals scored in August, October, and November is twice higher in contrast with Liverpool FC results. Moreover, the tendency line of goals scored away (figure 3) by Manchester City is more stable and above 2 goals.

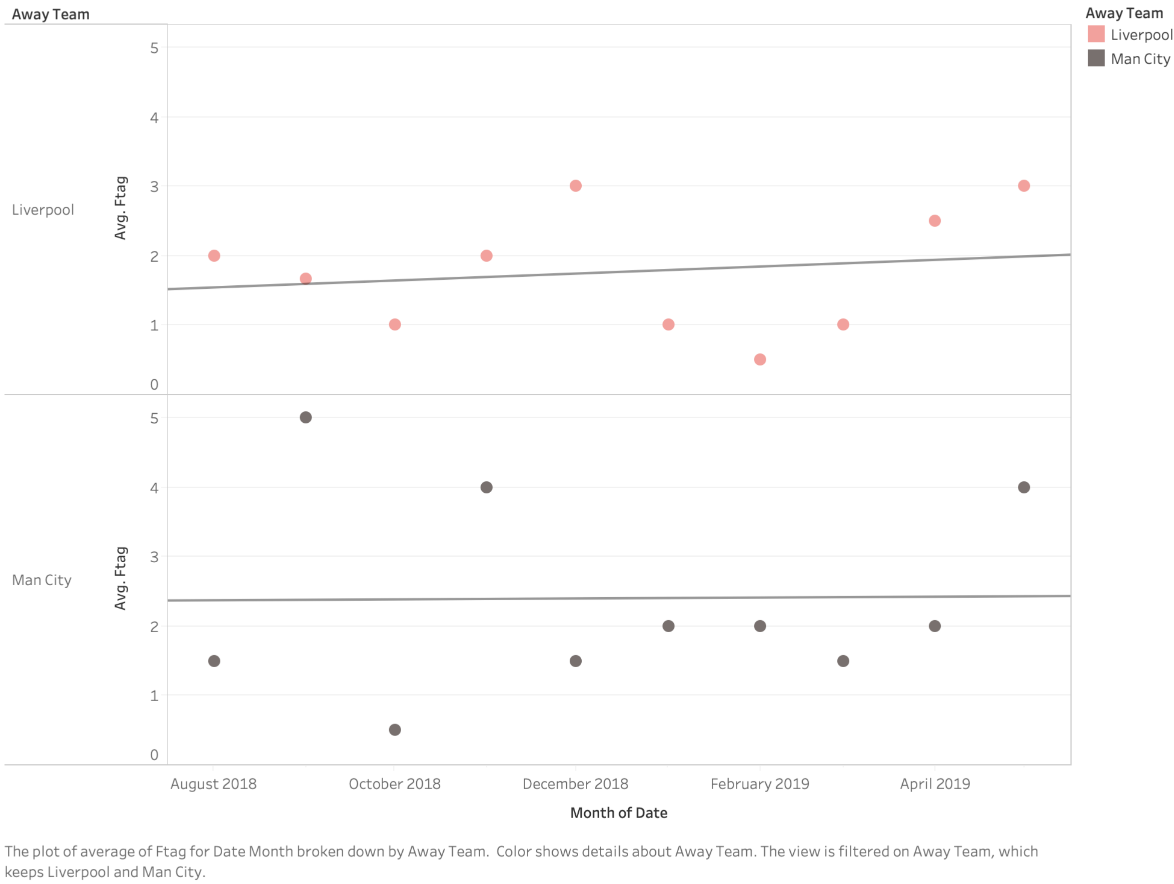


Figure 3. The average number of goals scored Away (Liverpool FC vs. Man City)

The graphs of the conceded goals at home (figure 4) by two teams contain two converse tendencies. Now Liverpool had a better performance at home at the beginning of the season. Two first months they didn't concede any goals. In the first half of the season (until January), Liverpool FC conceded less than Manchester City but after that, they significantly decreased their result. The reason why both clubs are conceded more goals in December in January is "Boxing Day." During the Boxing Day, soccer clubs play 2 games per weak. In this period every team is vulnerable to defeat because of high physical stress on players. Table 5 shows that Manchester City played in defense better than Liverpool at away games, and I think a high level of conceded goal by Liverpool in the last month was a crucial moment on championship rally.

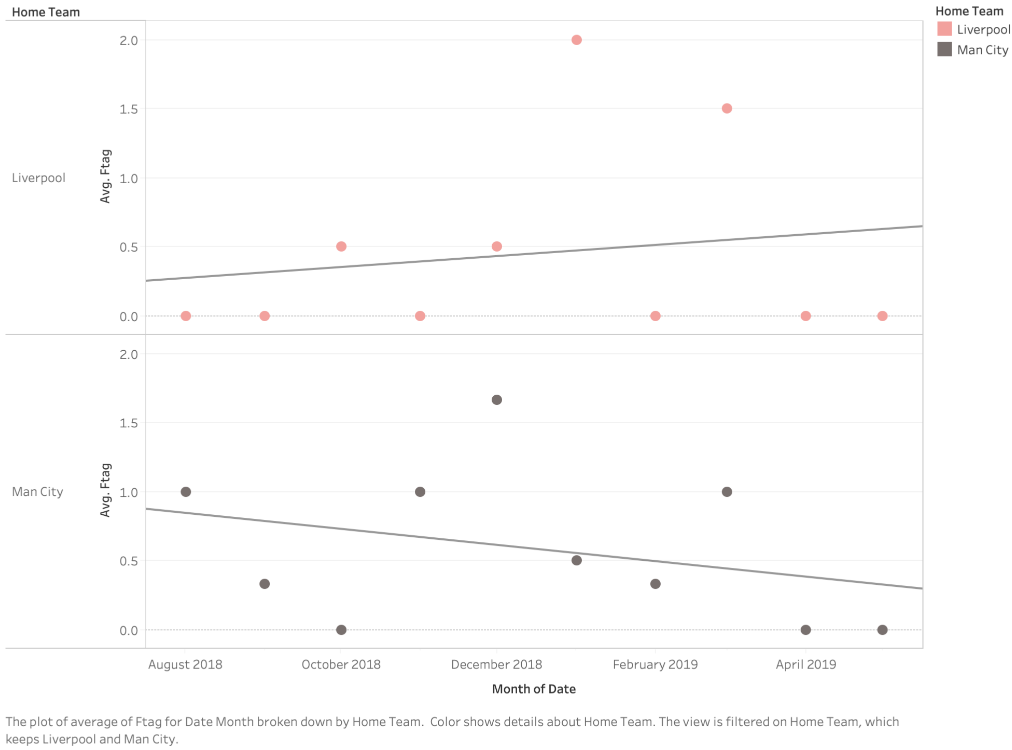


Figure 4. The average number of goals conceded at Home (Liverpool FC vs. Man City)

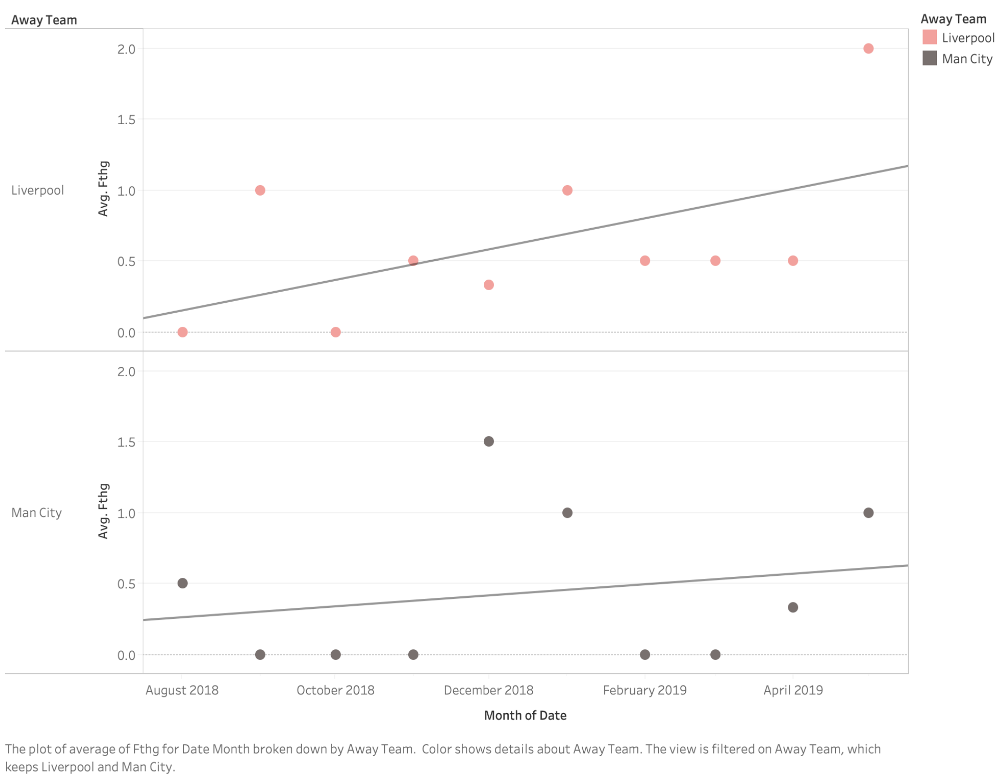


Figure 5. The average number of goals conceded Away (Liverpool FC vs. Man City)

Prediction model

The best way to predict the results of soccer matches is to use *Poisson distribution*. Poisson distribution is one of the earliest statistical methods of forecasting sports events because it is discrete probability distribution which can be used to model data that the number of events within a specific time period (e.g 90 minutes per game) with a known average rate of occurrence and independently of the time since the last event [3]. Our sports event fits the distribution conditions. I delete 10 last games to check actual results with the predicted results.

The regression model formula: $\boldsymbol{log(L) = mu + home + team_i + opponent_j}$

mu - the overall mean number of goals;

home - the effect on the number of goals a home team;

$team_i$ - the effect of team number *i*;

$opponent_j$ - the effect of team *j*.

Using the regression model in R

Fortunately, we have an appropriate *gml()* function in R which has already contained a Poisson regression model *(family=poisson())*. Nevertheless, we need to restructure the dataset and make it fittable in our regression model. To rearrange the dataset, we need to use the code script below (the whole script is presented in Appendix B). Table 2 shows how it should look like.

```
model <-  rbind(
 data.frame(Goals=epl$FTHG,
        Team=epl$HomeTeam,
        Opponent=epl$AwayTeam,
        Home=1),
 data.frame(Goals=epl$FTAG,
        Team=epl$AwayTeam,
        Opponent=epl$HomeTeam,
        Home=0))
```

#	Goals	Team	Opponent	Home
1	2	Man United	Leicester	1
2	2	Bournemouth	Cardiff	1
3	0	Fulham	Crystal Palace	1
4	0	Huddersfield	Chelsea	1
5	1	Newcastle	Tottenham	1
6	2	Watford	Brighton	1
7	2	Wolves	Everton	1
8	0	Arsenal	Man City	1
9	4	Liverpool	West Ham	1
10	0	Southampton	Burnley	1

Table 2. Restructured dataset

The next step is to use glm() function with the restructured dataset and create predicting model:

```
poisson_model <- glm(Goals ~ Home + Team + Opponent, family=poisson(link=log), data=model)
```

After that, I simulate the last 10 games by using the prediction model and compare it with the real results to check the accuracy of the Poisson model. According to Tables 3 and 4, if we consider only the outcome of the match (win, draw, lose), half of the predicted results match with actual results. Only one predicted match has the same score (Liverpool 2-0 Wolves).

#	HomeTeam	AwayTeam	home_goals	away_goals
1	Brighton	Man City	0	2
2	Burnley	Arsenal	1	2
3	Crystal Palace	Bournemouth	2	1
4	Fulham	Newcastle	1	1
5	Leicester	Chelsea	1	1
6	Liverpool	Wolves	2	0
7	Man United	Cardiff	3	1
8	Southampton	Huddersfield	2	1
9	Tottenham	Everton	2	1
10	Watford	West Ham	2	1

Table 3. Predicted results

#	HomeTeam	AwayTeam	Home_goals	Away_goals
1	Brighton	Man City	1	4
2	Burnley	Arsenal	1	3
3	Crystal Palace	Bournemouth	5	3
4	Fulham	Newcastle	0	4
5	Leicester	Chelsea	0	0
6	Liverpool	Wolves	2	0
7	Man United	Cardiff	0	2
8	Southampton	Huddersfield	1	1
9	Tottenham	Everton	2	2
10	Watford	West Ham	1	4

Table 4. Actual results

Conclusion

Sum up, the summary statistics give some enlightenment on the champion's winning. Manchester City soccer club has almost the best characteristics in all categories. They scored more and conceded less, which can imply they had both good attack and defense. Manchester City took advantage of the home stadium and scored goals as much as possible at the beginning of the championship. Moreover, a lot of conceded goals by Liverpool (2 goals conceded per match) in the last month might allow Manchester City to overtake them. According to the simulation of the last 10 games, the Poisson model based on the results of the 370 matches has 50% accuracy, which is fair enough. Nowadays, many factors could be affected by match results. For instance, possibly Manchester United lost their last game because they recently fired a head coach, etc. To make a prediction model more accurate, we need to include all relevant factors related to the particular soccer match.

References

1. "Premier League, Season 2018/2019", Football-Data, 16 October 2019, https://www.football-data.co.uk/mmz4281/1819/E0.csv
2. Tables, Premier League, 2019, https://www.premierleague.com/tables?co=1&se=210&ha=-1
3. Poisson distribution, Wikipedia, 23 November 2019, https://en.wikipedia.org/wiki/Poisson_distribution

Appendix A

Notes for Football Data

All data is in csv format, ready for use within standard spreadsheet applications. Please note that some abbreviations are no longer in use (in particular odds from specific bookmakers no longer used) and refer to data collected in earlier seasons. For a current list of what bookmakers are included in the dataset please visit http://www.football-data.co.uk/matches.php

Key to results data:

Div = League Division
Date = Match Date (dd/mm/yy)
Time = Time of match kick off
HomeTeam = Home Team
AwayTeam = Away Team
FTHG and HG = Full Time Home Team Goals
FTAG and AG = Full Time Away Team Goals
FTR and Res = Full Time Result (H=Home Win, D=Draw, A=Away Win)
HTHG = Half Time Home Team Goals
HTAG = Half Time Away Team Goals
HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)

Match Statistics (where available)
Attendance = Crowd Attendance
Referee = Match Referee
HS = Home Team Shots
AS = Away Team Shots
HST = Home Team Shots on Target
AST = Away Team Shots on Target
HHW = Home Team Hit Woodwork
AHW = Away Team Hit Woodwork
HC = Home Team Corners
AC = Away Team Corners
HF = Home Team Fouls Committed
AF = Away Team Fouls Committed
HFKC = Home Team Free Kicks Conceded
AFKC = Away Team Free Kicks Conceded
HO = Home Team Offsides
AO = Away Team Offsides
HY = Home Team Yellow Cards

AY = Away Team Yellow Cards
HR = Home Team Red Cards
AR = Away Team Red Cards
HBP = Home Team Bookings Points (10 = yellow, 25 = red)
ABP = Away Team Bookings Points (10 = yellow, 25 = red)

Note that Free Kicks Conceeded includes fouls, offsides and any other offense commmitted and will always be equal to or higher than the number of fouls. Fouls make up the vast majority of Free Kicks Conceded. Free Kicks Conceded are shown when specific data on Fouls are not available (France 2nd, Belgium 1st and Greece 1st divisions).

Note also that English and Scottish yellow cards do not include the initial yellow card when a second is shown to a player converting it into a red, but this is included as a yellow (plus red) for European games.

Key to 1X2 (match) betting odds data:

B365H = Bet365 home win odds
B365D = Bet365 draw odds
B365A = Bet365 away win odds
BSH = Blue Square home win odds
BSD = Blue Square draw odds
BSA = Blue Square away win odds
BWH = Bet&Win home win odds
BWD = Bet&Win draw odds
BWA = Bet&Win away win odds
GBH = Gamebookers home win odds
GBD = Gamebookers draw odds
GBA = Gamebookers away win odds
IWH = Interwetten home win odds
IWD = Interwetten draw odds
IWA = Interwetten away win odds
LBH = Ladbrokes home win odds
LBD = Ladbrokes draw odds
LBA = Ladbrokes away win odds
PSH and PH = Pinnacle home win odds
PSD and PD = Pinnacle draw odds
PSA and PA = Pinnacle away win odds
SOH = Sporting Odds home win odds
SOD = Sporting Odds draw odds

SOA = Sporting Odds away win odds
SBH = Sportingbet home win odds
SBD = Sportingbet draw odds
SBA = Sportingbet away win odds
SJH = Stan James home win odds
SJD = Stan James draw odds
SJA = Stan James away win odds
SYH = Stanleybet home win odds
SYD = Stanleybet draw odds
SYA = Stanleybet away win odds
VCH = VC Bet home win odds
VCD = VC Bet draw odds
VCA = VC Bet away win odds
WHH = William Hill home win odds
WHD = William Hill draw odds
WHA = William Hill away win odds

Bb1X2 = Number of BetBrain bookmakers used to calculate match odds averages and maximums
BbMxH = Betbrain maximum home win odds
BbAvH = Betbrain average home win odds
BbMxD = Betbrain maximum draw odds
BbAvD = Betbrain average draw win odds
BbMxA = Betbrain maximum away win odds
BbAvA = Betbrain average away win odds

MaxH = Market maximum home win odds
MaxD = Market maximum draw win odds
MaxA = Market maximum away win odds
AvgH = Market average home win odds
AvgD = Market average draw win odds
AvgA = Market average away win odds

Key to total goals betting odds:

BbOU = Number of BetBrain bookmakers used to calculate over/under 2.5 goals (total goals) averages and maximums
BbMx>2.5 = Betbrain maximum over 2.5 goals
BbAv>2.5 = Betbrain average over 2.5 goals
BbMx<2.5 = Betbrain maximum under 2.5 goals

BbAv<2.5 = Betbrain average under 2.5 goals

GB>2.5 = Gamebookers over 2.5 goals
GB<2.5 = Gamebookers under 2.5 goals
B365>2.5 = Bet365 over 2.5 goals
B365<2.5 = Bet365 under 2.5 goals
P>2.5 = Pinnacle over 2.5 goals
P<2.5 = Pinnacle under 2.5 goals
Max>2.5 = Market maximum over 2.5 goals
Max<2.5 = Market maximum under 2.5 goals
Avg>2.5 = Market average over 2.5 goals
Avg<2.5 = Market average under 2.5 goals

Key to Asian handicap betting odds:

BbAH = Number of BetBrain bookmakers used to Asian handicap averages and maximums
BbAHh = Betbrain size of handicap (home team)
AHh = Market size of handicap (home team) (since 2019/2020)
BbMxAHH = Betbrain maximum Asian handicap home team odds
BbAvAHH = Betbrain average Asian handicap home team odds
BbMxAHA = Betbrain maximum Asian handicap away team odds
BbAvAHA = Betbrain average Asian handicap away team odds

GBAHH = Gamebookers Asian handicap home team odds
GBAHA = Gamebookers Asian handicap away team odds
GBAH = Gamebookers size of handicap (home team)
LBAHH = Ladbrokes Asian handicap home team odds
LBAHA = Ladbrokes Asian handicap away team odds
LBAH = Ladbrokes size of handicap (home team)
B365AHH = Bet365 Asian handicap home team odds
B365AHA = Bet365 Asian handicap away team odds
B365AH = Bet365 size of handicap (home team)
PAHH = Pinnacle Asian handicap home team odds
PAHA = Pinnacle Asian handicap away team odds
MaxAHH = Market maximum Asian handicap home team odds
MaxAHA = Market maximum Asian handicap away team odds
AvgAHH = Market average Asian handicap home team odds
AvgAHA = Market average Asian handicap away team odds

Closing odds (last odds before match starts)

As above but with an additional "C" character following the bookmaker abbreviation/Max/Avg

Football-Data would like to acknowledge the following sources which have been utilised in the compilation of Football-Data's results and odds files.

Current results (full time, half time)
Xcores - http://www.xcores .com

Match statistics
BBC, ESPN Soccer, Bundesliga.de, Gazzetta.it and Football.fr

Bookmakers betting odds
Individual bookmakers

Betting odds for weekend games are collected Friday afternoons, and on Tuesday afternoons for midweek games.

Additional match statistics (corners, shots, bookings, referee etc.) for the 2000/01 and 2001/02 seasons for the English, Scottish and German leagues were provided by Sports.com (now under new ownership and no longer available).

Appendix B

```
# load library
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
library(tidyr)
library(skellam)

# set word directory
setwd("~/Dropbox/INTO George Mason University/AIT580 Analytics Big Data to In
formation/Final Project")

# load data
data <- read.csv("Premier League Season 2018-2019.csv")

# remove last 10 games to compare results with model
epl <- head(data,-10)
str(epl)

## 'data.frame':    370 obs. of  62 variables:
##  $ Div       : Factor w/ 1 level "E0": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Date      : Factor w/ 108 levels "01/01/2019","01/04/2019",..: 40 44 44
44 44 44 44 49 49 49 ...
##  $ HomeTeam  : Factor w/ 20 levels "Arsenal","Bournemouth",..: 14 2 9 10 1
5 18 20 1 12 16 ...
##  $ AwayTeam  : Factor w/ 20 levels "Arsenal","Bournemouth",..: 11 5 7 6 17
3 8 13 19 4 ...
##  $ FTHG      : int  2 2 0 0 1 2 2 0 4 0 ...
##  $ FTAG      : int  1 0 2 3 2 0 2 2 0 0 ...
##  $ FTR       : Factor w/ 3 levels "A","D","H": 3 3 1 1 1 3 2 1 3 2 ...
##  $ HTHG      : int  1 1 0 0 1 1 1 0 2 0 ...
##  $ HTAG      : int  0 0 1 2 2 0 1 1 0 0 ...
##  $ HTR       : Factor w/ 3 levels "A","D","H": 3 3 1 1 1 3 2 1 3 2 ...
##  $ Referee   : Factor w/ 18 levels "A Madley","A Marriner",..: 2 9 13 4 12
8 5 14 3 7 ...
##  $ HS        : int  8 12 15 6 15 19 11 9 18 18 ...
##  $ AS        : int  13 10 10 13 15 6 6 17 5 16 ...
##  $ HST       : int  6 4 6 1 2 5 4 3 8 3 ...
##  $ AST       : int  4 1 9 4 5 0 5 8 2 6 ...
##  $ HF        : int  11 11 9 9 11 10 8 11 14 10 ...
```

```
##  $ AF        : int  8 9 11 8 12 16 7 14 9 9 ...
##  $ HC        : int  2 7 5 2 3 8 3 2 5 8 ...
##  $ AC        : int  5 4 5 5 5 2 6 9 4 5 ...
##  $ HY        : int  2 1 1 2 2 2 0 2 1 0 ...
##  $ AY        : int  1 1 2 1 2 2 1 2 2 1 ...
##  $ HR        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AR        : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ B365H     : num  1.57 1.9 2.5 6.5 3.9 2.37 2.37 4 1.25 1.85 ...
##  $ B365D     : num  3.9 3.6 3.4 4 3.5 3.2 3.3 3.8 6.5 3.5 ...
##  $ B365A     : num  7.5 4.5 3 1.61 2.04 3.4 3.3 1.95 14 5 ...
##  $ BWH       : num  1.53 1.9 2.45 6.25 3.8 2.35 2.35 3.7 1.2 1.8 ...
##  $ BWD       : num  4 3.4 3.3 3.9 3.5 3.1 3.2 3.75 6.75 3.5 ...
##  $ BWA       : num  7.5 4.4 2.95 1.57 2 3.3 3.2 1.95 14 4.75 ...
##  $ IWH       : num  1.55 1.9 2.4 6.2 3.7 2.2 2.25 3.6 1.25 1.8 ...
##  $ IWD       : num  3.8 3.5 3.3 4 3.35 3.3 3.35 3.6 6.1 3.6 ...
##  $ IWA       : num  7 4.1 2.95 1.55 2.05 3.4 3.2 2 11 4.5 ...
##  $ PSH       : num  1.58 1.89 2.5 6.41 3.83 2.43 2.36 4 1.27 1.86 ...
##  $ PSD       : num  3.93 3.63 3.46 4.02 3.57 3.22 3.4 3.97 6.35 3.51 ...
##  $ PSA       : num  7.5 4.58 3 1.62 2.08 ...
##  $ WHH       : num  1.57 1.91 2.45 5.8 3.8 2.38 2.3 3.8 1.25 1.83 ...
##  $ WHD       : num  3.8 3.5 3.3 3.9 3.2 3 3.2 3.8 5.5 3.25 ...
##  $ WHA       : num  6 4 2.8 1.57 2.05 3.3 3.2 1.91 12 4.8 ...
##  $ VCH       : num  1.57 1.87 2.5 6.5 3.9 2.4 2.38 3.9 1.25 1.85 ...
##  $ VCD       : num  4 3.6 3.4 4 3.4 3.2 3.3 4 6.5 3.4 ...
##  $ VCA       : num  7 4.75 3 1.62 2.1 3.4 3.3 1.91 13 5.2 ...
##  $ Bb1X2     : int  39 39 39 38 39 39 38 39 38 39 ...
##  $ BbMxH     : num  1.6 1.93 2.6 6.85 4.01 2.48 2.41 4.15 1.29 1.9 ...
##  $ BbAvH     : num  1.56 1.88 2.47 6.09 3.83 2.36 2.33 3.83 1.25 1.84 ...
##  $ BbMxD     : num  4.2 3.71 3.49 4.07 3.57 3.3 3.4 4 6.79 3.61 ...
##  $ BbAvD     : num  3.92 3.53 3.35 3.9 3.4 3.14 3.27 3.8 6.22 3.43 ...
##  $ BbMxA     : num  8.05 4.75 3.05 1.66 2.12 3.42 3.4 2 15 5.2 ...
##  $ BbAvA     : num  7.06 4.37 2.92 1.61 2.05 3.31 3.23 1.92 12.3 4.8 ...
##  $ BbOU      : int  38 38 38 37 38 37 36 36 33 37 ...
##  $ BbMx.2.5  : num  2.12 2.05 2 2.05 2.1 2.46 2.2 1.6 1.49 2.45 ...
##  $ BbAv.2.5  : num  2.03 1.98 1.95 1.98 2.01 2.35 2.09 1.55 1.44 2.34 ...
##  $ BbMx.2.5.1: num  1.85 1.92 1.96 1.9 1.88 1.67 1.83 2.55 2.88 1.67 ...
##  $ BbAv.2.5.1: num  1.79 1.83 1.87 1.84 1.81 1.59 1.75 2.42 2.72 1.6 ...
##  $ BbAH      : int  17 20 22 23 20 22 22 20 21 20 ...
##  $ BbAHh     : num  -0.75 -0.75 -0.25 1 0.25 -0.25 -0.25 0.75 -1.75 -0.75
...
##  $ BbMxAHH   : num  1.75 2.2 2.18 1.84 2.2 2.07 2.04 1.78 1.95 2.19 ...
##  $ BbAvAHH   : num  1.7 2.13 2.11 1.8 2.12 2.01 1.98 1.74 1.9 2.11 ...
##  $ BbMxAHA   : num  2.29 1.8 1.81 2.13 1.8 1.9 1.92 2.21 2.06 1.82 ...
##  $ BbAvAHA   : num  2.21 1.75 1.77 2.06 1.76 1.86 1.88 2.15 1.97 1.76 ...
##  $ PSCH      : num  1.55 1.88 2.62 7.24 4.74 2.58 2.44 4.43 1.25 2.03 ...
##  $ PSCD      : num  4.07 3.61 3.38 3.95 3.53 3.08 3.23 4.13 6.95 3.19 ...
##  $ PSCA      : num  7.69 4.7 2.9 1.58 1.89 3.22 3.32 1.81 12 4.65 ...

# build data frame for poisson model
model <-  rbind(
```

```r
  data.frame(Goals=epl$FTHG,
             Team=epl$HomeTeam,
             Opponent=epl$AwayTeam,
             Home=1),
  data.frame(Goals=epl$FTAG,
             Team=epl$AwayTeam,
             Opponent=epl$HomeTeam,
             Home=0))
head(model,10)
```

```
##    Goals         Team       Opponent Home
## 1      2   Man United      Leicester    1
## 2      2  Bournemouth        Cardiff    1
## 3      0       Fulham Crystal Palace    1
## 4      0 Huddersfield        Chelsea    1
## 5      1    Newcastle      Tottenham    1
## 6      2      Watford       Brighton    1
## 7      2       Wolves        Everton    1
## 8      0      Arsenal       Man City    1
## 9      4    Liverpool       West Ham    1
## 10     0  Southampton        Burnley    1
```

```r
# fit model and get a summary
poisson_model <- glm(Goals ~ Home + Team + Opponent, family=poisson(link=log)
, data=model)
summary(poisson_model)
```

```
##
## Call:
## glm(formula = Goals ~ Home + Team + Opponent, family = poisson(link = log)
,
##     data = model)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.17511  -1.02684  -0.05616   0.48462   2.98913
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               0.49255    0.19153   2.572 0.010120 *
## Home                      0.25265    0.06268   4.031 5.56e-05 ***
## TeamBournemouth          -0.27027    0.18251  -1.481 0.138637
## TeamBrighton             -0.73344    0.20938  -3.503 0.000460 ***
## TeamBurnley              -0.44926    0.19248  -2.334 0.019596 *
## TeamCardiff              -0.77017    0.21376  -3.603 0.000315 ***
## TeamChelsea              -0.12488    0.17402  -0.718 0.472976
## TeamCrystal Palace       -0.41252    0.19021  -2.169 0.030099 *
## TeamEverton              -0.31651    0.18342  -1.726 0.084416 .
## TeamFulham               -0.69864    0.20946  -3.335 0.000852 ***
## TeamHuddersfield         -1.18201    0.24914  -4.744 2.09e-06 ***
```

```
## TeamLeicester          -0.32612    0.18448  -1.768 0.077101 .
## TeamLiverpool           0.18490    0.16091   1.149 0.250511
## TeamMan City            0.22907    0.15931   1.438 0.150462
## TeamMan United         -0.06212    0.17271  -0.360 0.719084
## TeamNewcastle          -0.61039    0.20184  -3.024 0.002493 **
## TeamSouthampton        -0.43869    0.19284  -2.275 0.022912 *
## TeamTottenham          -0.09173    0.17263  -0.531 0.595136
## TeamWatford            -0.31059    0.18452  -1.683 0.092338 .
## TeamWest Ham           -0.37878    0.18778  -2.017 0.043679 *
## TeamWolves             -0.42390    0.18891  -2.244 0.024834 *
## OpponentBournemouth     0.24578    0.18857   1.303 0.192441
## OpponentBrighton        0.08898    0.19497   0.456 0.648101
## OpponentBurnley         0.24133    0.18821   1.282 0.199748
## OpponentCardiff         0.29496    0.18614   1.585 0.113057
## OpponentChelsea        -0.25371    0.21407  -1.185 0.235938
## OpponentCrystal Palace -0.02685    0.20040  -0.134 0.893410
## OpponentEverton        -0.13655    0.20713  -0.659 0.509733
## OpponentFulham          0.38609    0.18201   2.121 0.033901 *
## OpponentHuddersfield    0.35513    0.18294   1.941 0.052224 .
## OpponentLeicester      -0.05948    0.20249  -0.294 0.768947
## OpponentLiverpool      -0.81119    0.25624  -3.166 0.001547 **
## OpponentMan City       -0.80828    0.25625  -3.154 0.001609 **
## OpponentMan United      0.02311    0.19852   0.116 0.907343
## OpponentNewcastle      -0.08018    0.20244  -0.396 0.692067
## OpponentSouthampton     0.20434    0.18914   1.080 0.279984
## OpponentTottenham      -0.31066    0.21727  -1.430 0.152772
## OpponentWatford         0.07114    0.19582   0.363 0.716397
## OpponentWest Ham        0.05754    0.19669   0.293 0.769854
## OpponentWolves         -0.12989    0.20714  -0.627 0.530607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 942.63  on 739  degrees of freedom
## Residual deviance: 732.30  on 700  degrees of freedom
## AIC: 2140.2
##
## Number of Fisher Scoring iterations: 5

#Simulate last gameweek
last_gameweek <- data[371:380, 3:4]

home_goals <- c(predict(poisson_model, data.frame(Home=1, Team="Brighton", Op
ponent="Man City"), type="response"),
predict(poisson_model, data.frame(Home=1, Team="Burnley", Opponent="Arsenal")
, type="response"),
predict(poisson_model, data.frame(Home=1, Team="Crystal Palace", Opponent="Bo
urnemouth"), type="response"),
predict(poisson_model, data.frame(Home=1, Team="Fulham", Opponent="Newcastle"
```

```
), type="response"),
predict(poisson_model, data.frame(Home=1, Team="Leicester", Opponent="Chelsea
"), type="response"),
predict(poisson_model, data.frame(Home=1, Team="Liverpool", Opponent="Wolves"
), type="response"),
predict(poisson_model, data.frame(Home=1, Team="Man United", Opponent="Cardif
f"), type="response"),
predict(poisson_model, data.frame(Home=1, Team="Southampton", Opponent="Hudde
rsfield"), type="response"),
predict(poisson_model, data.frame(Home=1, Team="Tottenham", Opponent="Everton
"), type="response"),
predict(poisson_model, data.frame(Home=1, Team="Watford", Opponent="West Ham"
), type="response"))


away_goals <- c(predict(poisson_model, data.frame(Home=0, Team="Man City", Op
ponent="Brighton"), type="response"),
predict(poisson_model, data.frame(Home=0, Team="Arsenal", Opponent="Burnley")
, type="response"),
predict(poisson_model, data.frame(Home=0, Team="Bournemouth", Opponent="Cryst
al Palace"), type="response"),
predict(poisson_model, data.frame(Home=0, Team="Newcastle", Opponent="Fulham"
), type="response"),
predict(poisson_model, data.frame(Home=0, Team="Chelsea", Opponent="Leicester
"), type="response"),
predict(poisson_model, data.frame(Home=0, Team="Wolves", Opponent="Liverpool"
), type="response"),
predict(poisson_model, data.frame(Home=0, Team="Cardiff", Opponent="Man Unite
d"), type="response"),
predict(poisson_model, data.frame(Home=0, Team="Huddersfield", Opponent="Sout
hampton"), type="response"),
predict(poisson_model, data.frame(Home=0, Team="Everton", Opponent="Tottenham
"), type="response"),
predict(poisson_model, data.frame(Home=0, Team="West Ham", Opponent="Watford"
), type="response"))

last_gameweek <- last_gameweek %>% mutate(round(home_goals,0), round(away_goa
ls,0))
last_gameweek

##          HomeTeam     AwayTeam round(home_goals, 0) round(away_goals, 0)
## 1        Brighton     Man City                    0                    2
## 2         Burnley      Arsenal                    1                    2
## 3  Crystal Palace  Bournemouth                    2                    1
## 4          Fulham    Newcastle                    1                    1
## 5       Leicester      Chelsea                    1                    1
## 6       Liverpool       Wolves                    2                    0
## 7      Man United      Cardiff                    3                    1
## 8     Southampton Huddersfield                    2                    1
```

```
## 9       Tottenham      Everton                    2                    1
## 10        Watford     West Ham                    2                    1
```