The dataset I am using to analyze is Premier League match from 1993-1994 to 2017-2018. It also includes half time results, but only from 1995-96 to 2017-18. Columns include Division (denoted as E0), HomeTeam, AwayTeam, FTHG (final time home goals), FTAG (final time away goals), FTR (full time result), HTHG (half time home goals), HTAG (half time away goals), HTR (half time result), and season.

Some other variables I used from the csv file with betting rates:

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

My project will contain couple of parts:

1. Analyzing some data of the league，mostly about ELP is a popular spectator league, such as total goals for each match, the overall standing historically.
   1. Fixing the data

I fixed the data where has NA, for 93 94 season there are no half time score and result, So I fix them by using the total score divided by two and same result as full time result

* 1. Add a column Total goals for each match
  2. Add a column Goal difference each match
  3. Count outliers of those two matchings, which shows how many game has big score, and how many has the game is dominated by one side
  4. Create a new table with Total goals, home goals, away goals in each season (show by using histogram)
  5. Create a table with goals in each year, which shows in each season there will be around 1000 goals appear.
  6. Home team goals vs away team goals
     1. This shows that home team usually perform better.
  7. Creating a new database historical standings
     1. This shows MU and Arsenal are the best two team in past 15 years.
     2. Creating a histogram x is team, y is the points that they get historically
  8. Calculating the correlation between team score and team being scored by their opponent. The result is that score the other team and being scored have weak negative relationship, both showed by Pearson and Spearman correlation.

1. using query to output the historical record of one team vs another.
   1. Creating a database, import the updated data
   2. Output a game result on a date
   3. Output a historical record between two teams
2. Making a prediction for this current season 2018-2019, by build couple of models and use the previous data as train set, compare with couple of matches this season already done to check if the model accuracy.

Data compiled into one file from this site:

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

1. Adding new dataset with betting rates into the database.
2. Pick the useful betting rates. The betting companies’ names were shown above
3. Normalize the number using z-score standardization
4. Fixing the column has same name for both dataset in 2017 and 2018
5. Adding dummy variables to compute the win, draw, lose (index W=2, D =1, L=0)
6. Create trainset and test set.
7. Use class package knn make prediction
8. Guess one game, in this season with providing some betting rates in a data frame.
9. Then guess the test set
10. Testing accuracy and show confusion matrix, and the accuracy is around 95% above usually\
11. Use linear regression model
12. Same steps 9 to 10 as above
13. The Accuracy is usually around 45%
14. Find the reason for linear model is not correct is that game does not have a continuous relationship, so linear regression model is not appropriate, knn is better.
15. Using knn predict this season as well repeat steps 9 and 10.