**IE 582**

**STATISTICAL LEARNING FOR DATA MINING**

**Homework 1**

By

Gizem KURTBAY

**Course Instructor** : Mustafa Gökçe BAYDOĞAN

**Submitted to** : İpek DURSUN

**Date of Submission** : 18.10.2018

**Department of Industrial Engineering**

**Boğaziçi University**

**Bebek, Istanbul**

**Task 1 a**

The goodness of the over/under decisions of 5 different bookmakers is checked in this task by using scatter plots. Aline with the equation of x=y is also plotted as a red line to give a better visualization to the plots. The data are expected to be around the x=y line if the bookmaker’s decisions are good. Initial and final odds are found and plotted against the results for all bookmakers. The code is given in the appendix with all required information and explanations.

Plot of empirical data versus the initial probabilities of Pinnacle is plotted in the figure below. As it can be seen from the figure, most of the data is around the x=y line which show that Pinnacle had good estimations about the results. There are outliers in the graph, however unexpected reasons may have affected the results of matches.

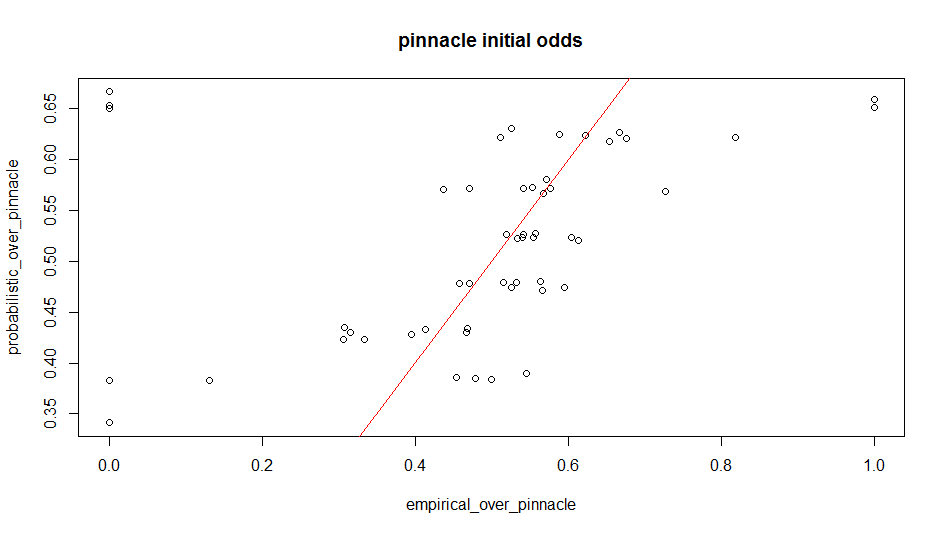


Figure 1: Pinnacle Initial Odds

Plot of empirical data versus the final probabilities of Pinnacle is plotted in the figure below. As it can be seen from the figure, most of the data is around the x=y line which show that Pinnacle had good estimations about the results. Moreover, the data points are closer to x=y line compared to the initial probabilities plot, which shows that the estimations got better through the time before the results are obtained. There are outliers in the graph, however unexpected reasons may have affected the results of matches even during the match time.

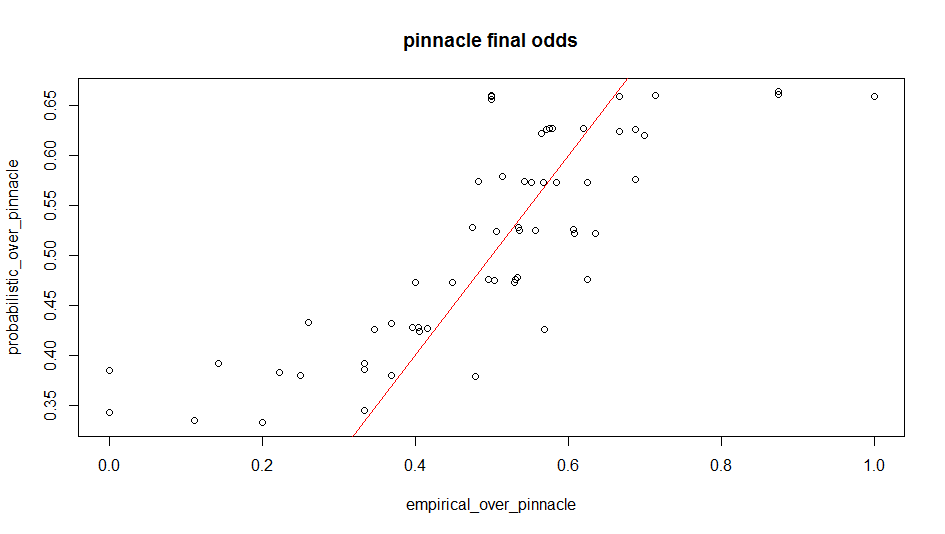


Figure 2: Pinnacle Final Odds

Plot of empirical data versus the initial probabilities of bet365 is plotted in the figure below. As it can be seen from the figure, most of the data is around the x=y line which show that bet365 had good estimations about the results. There are outliers in the graph, however unexpected reasons may have affected the results of matches, and the number of outliers is less than the number of outliers of Pinnacle’s, which shows bet365 may be more stable in case of unexpected situations. However, when we check the data points around the x=y line, the data of bet365 is spread wider compared to the data of Pinnacle. It can be observed that Pinnacle may guess closer to real results generally, however bet365 is more robust for unexpected reasons to cause outliers.

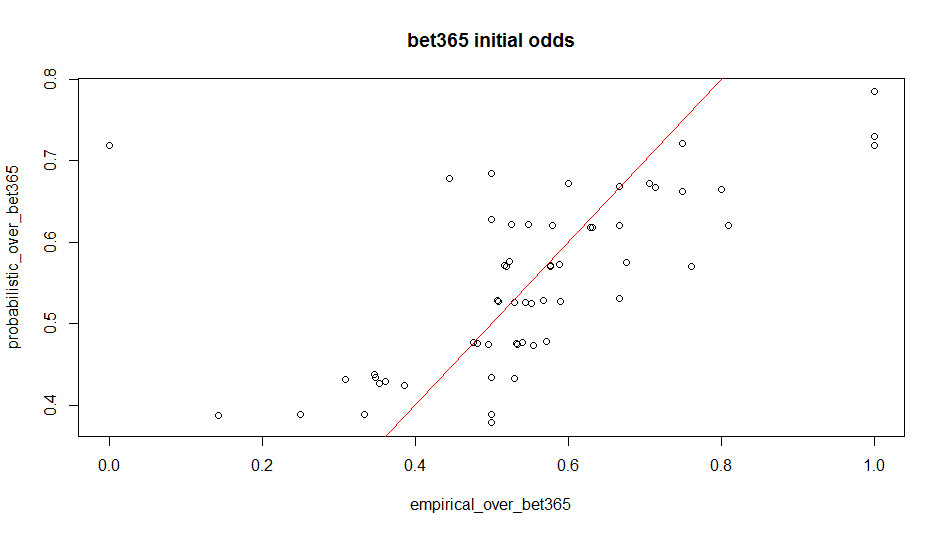


Figure 3: Bet365 Initial Odds

Plot of empirical data versus the final probabilities of bet 365 is plotted in the figure below. As it can be seen from the figure, most of the data is around the x=y line which show that bet365 had good estimations about the results. Locations of some of the data points changed around x=y line compared to the initial probabilities plot, and it can be seen that the data points are more homogeneous around the x=y plot. Initially the number of data points below x=y was higher compared to the number above the x=y line, and the numbers got closer to each other as time passed. There are some outliers in the graph, however unexpected reasons may have affected the results of matches even during the match time.

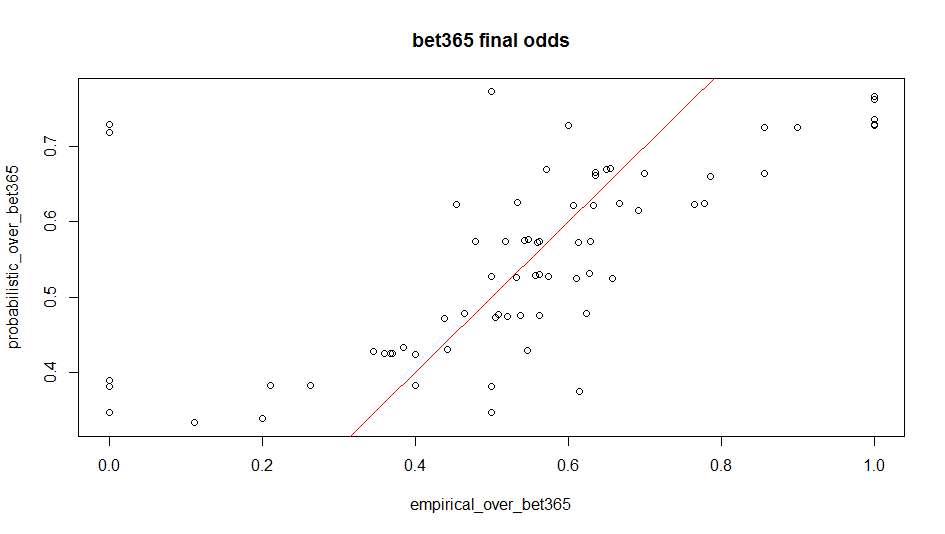


Figure 4: Bet365 Final Odds

Plot of empirical data versus the initial probabilities of betway is plotted in the figure below. As it can be seen from the figure, most of the data is around the x=y line which show that betway had good estimations about the results. There are outliers in the graph, however unexpected reasons may have affected the results of matches.

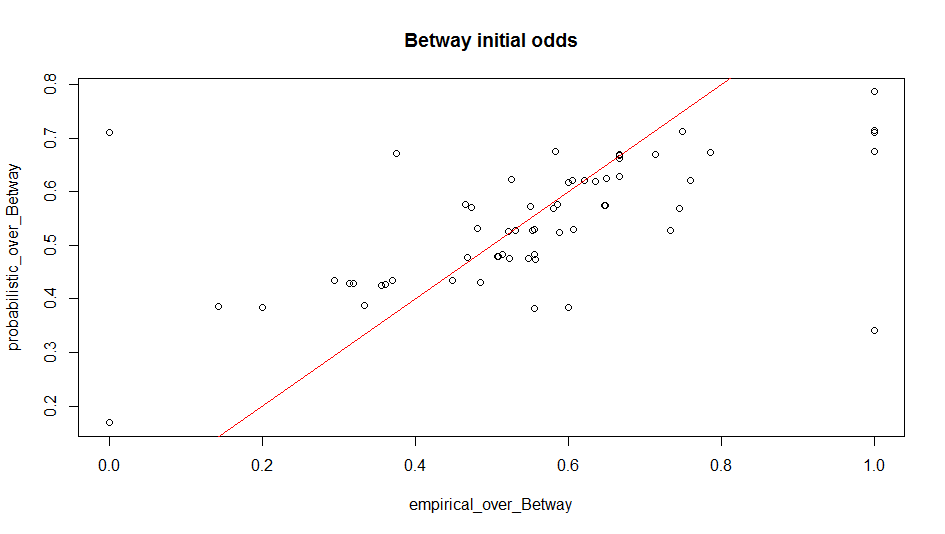


Figure 5: Betway Initial Odds

Plot of empirical data versus the final probabilities of betway is plotted in the figure below. As it can be seen from the figure, most of the data is still around the x=y line which show that betway had good estimations about the results. Even though, it seems like the data points are more far away from x=y line compared to the initial probabilities plot, it can be seen that this is not true if the scale of y axis’s are compared. It can be seen that the points around x=y axis stayed similar to their initial positions on the plot. Moreover it can be seen that some outliers are prevented and since the outlier around the location of 0 to 0.01 is gone during the time, the y axis scale got clearer to observe closer.

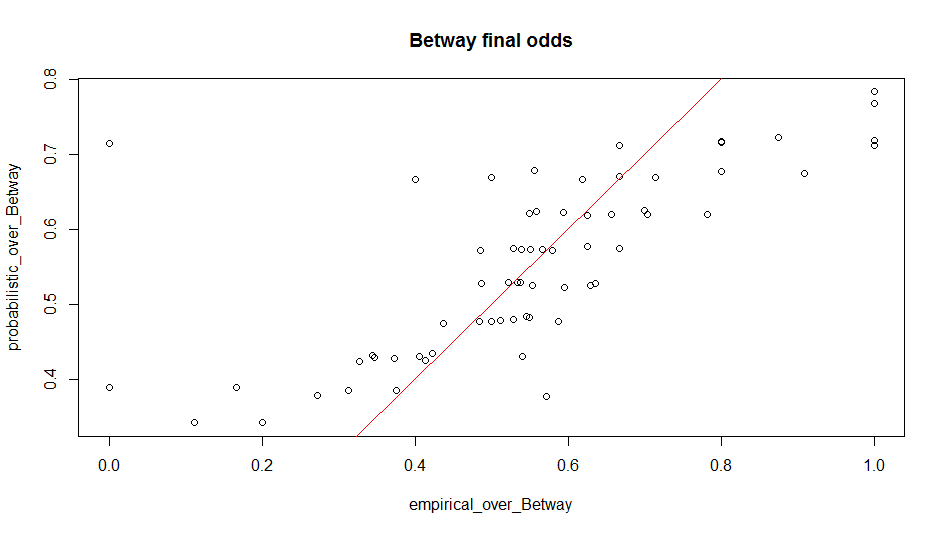


Figure 6: Betway Final Odds

Plot of empirical data versus the initial probabilities of bwin is plotted in the figure below. As it can be seen from the figure, most of the data is around the x=y line which show that bwin had good estimations about the results. There are some outliers in the graph, however unexpected reasons may have affected the results of matches.

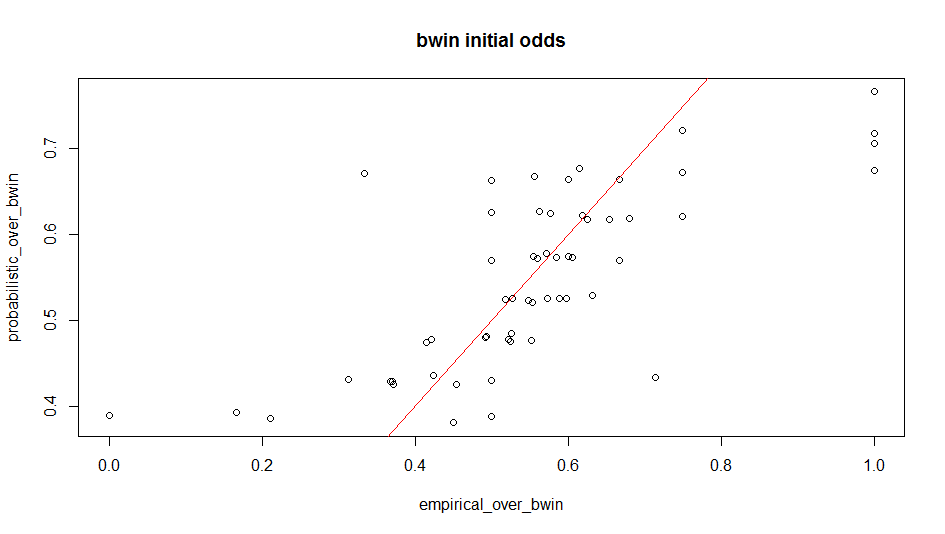


Figure 7: Bwin Initial Odds

Plot of empirical data versus the final probabilities of bwin is plotted in the figure below. As it can be seen from the figure, most of the data is around the x=y line which show that bwin had good estimations about the results. Locations of some of the data points changed around x=y line compared to the initial probabilities plot, and it can be seen that the data points are more homogeneous around the x=y plot. Initially the number of data points below x=y was higher compared to the number above the x=y line, and the numbers got closer to each other as time passed similar to the observation of bet365. There are some outliers in the graph, however unexpected reasons may have affected the results of matches even during the match time.

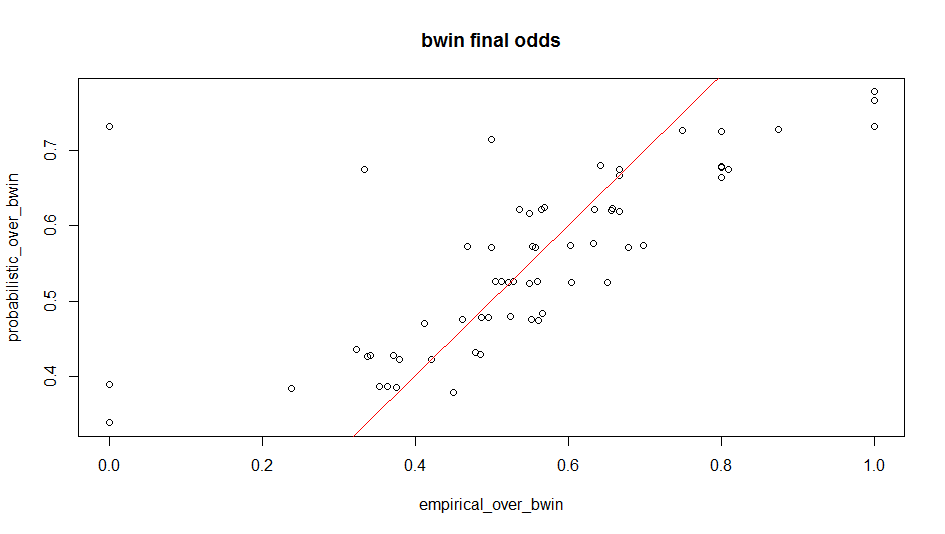


Figure 8: Bwin Final Odds

Plot of empirical data versus the initial probabilities of betsson is plotted in the figure below. As it can be seen from the figure, most of the data is around the x=y line which show that betsson had good estimations about the results. There are some outliers in the graph, however unexpected reasons may have affected the results of matches. The outliers are mostly around the empirical values of 0.00 and 1.0.

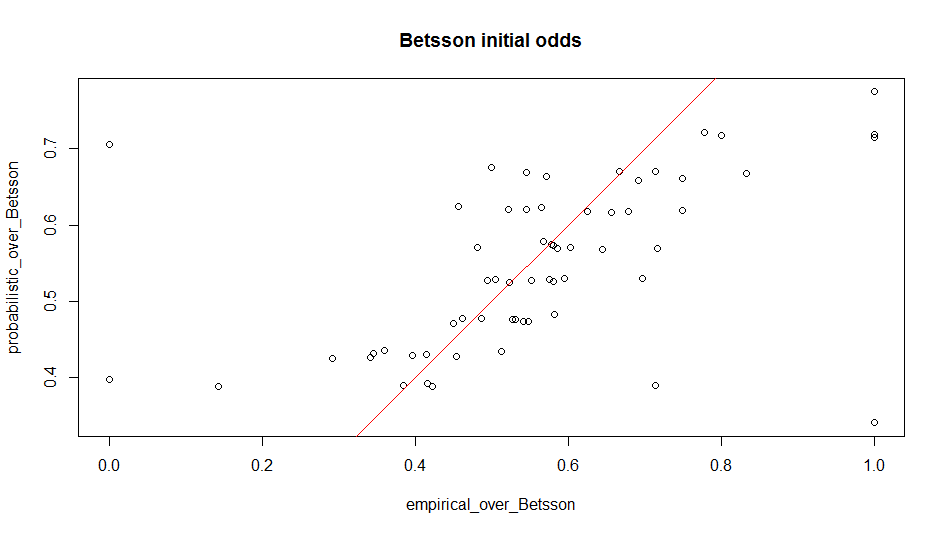


Figure 9: Betsson Initial Odds

Plot of empirical data versus the final probabilities of betsson is plotted in the figure below. As it can be seen from the figure, most of the data is around the x=y line which show that betsson had good estimations about the results. Locations of some of the data points changed around x=y line compared to the initial probabilities plot, and it can be seen that the data points are more homogeneous around the x=y plot. Initially the number of data points below x=y was higher compared to the number above the x=y line, and the numbers got closer to each other as time passed similar to the observation of bet365 and bwin. There are some outliers in the graph, however unexpected reasons may have affected the results of matches even during the match time.

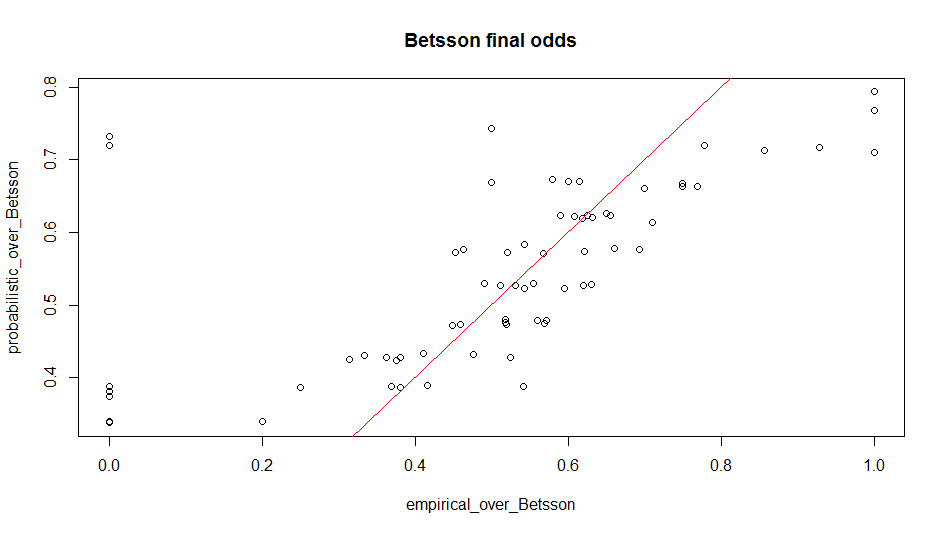


Figure 10: Betsson Final Odds

**Task 1 b**

Pinnacle is chosen as the bookmaker to check the constancy of determining over during time. Line plots are used for each bin in this task. The bin between 0.0 - 0.3 and 0.7 – 1.0 did not have enough data, therefore they could not be plotted for initial and final odds. Also 0.3 to 0.35 bin interval is plotted only for final odds comparison even though time comparison cannot be made without initial odds data. The code is given in the appendix with all required information and explanations.

It can be seen that, for the bin interval of 0.30 to 0.35 of final odds comparison, probabilities stay constant around the value of 0.35 while the empirical data fluctuates between 0 and 0.35. Expected probabilities are smoother than real results.

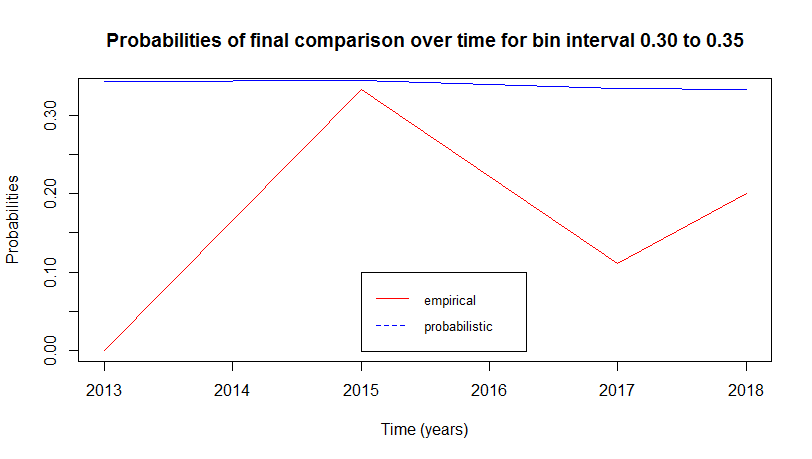


Figure 11: Final Probabilities for Bin Interval of 0.30 to 0.35 for Pinnacle

It can be seen that, for the bin interval of 0.35 to 0.40 of initial odds comparison, probabilities stay constant around the value of 0.4 while the empirical data fluctuates between 0 and 0.55. Expected probabilities are smoother than real results. There are empirical values both above and below the probabilistic values.

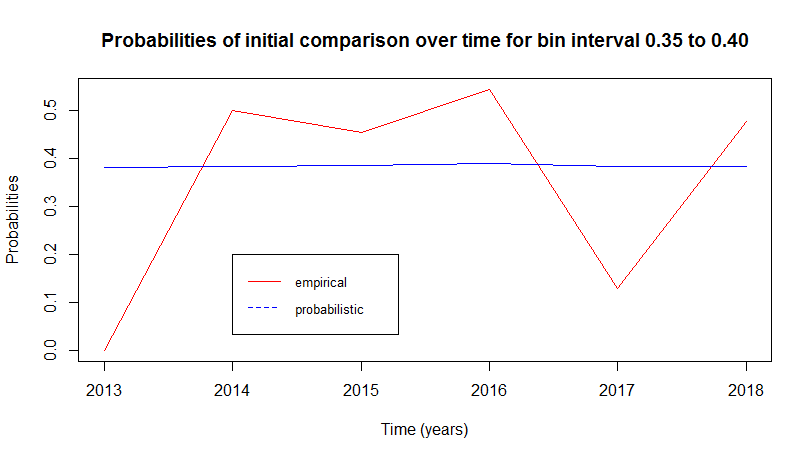


Figure 12: Initial Probabilities for Bin Interval of 0.35 to 0.40 for Pinnacle

It can be seen that, for the bin interval of 0.35 to 0.40 of final odds comparison, probabilities stay constant around the value of 0.4 while the empirical data fluctuates between 0 and 0.45. Expected probabilities are smoother than real results. There were empirical values both above and below the probabilistic values initially, however in final time, the empirical values are mostly below probabilistic values, which shows a possible effect of time on probabilities.

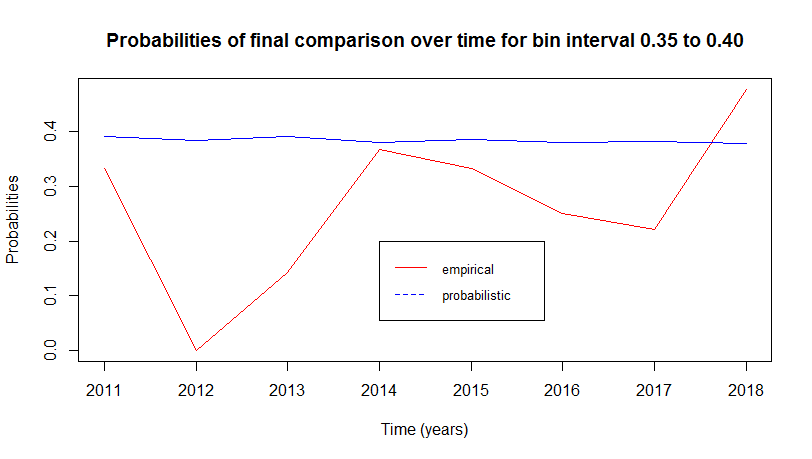


Figure 13: Final Probabilities for Bin Interval of 0.35 to 0.40 for Pinnacle

It can be seen that, for the bin interval of 0.4 to 0.45 of initial odds comparison, probabilities stay constant around the value of 0.44 while the empirical data fluctuates between 0.30 and 0.50. Expected probabilities are smoother than real results. There are empirical values both above and below the probabilistic values, but they are mostly below the probabilistic line.

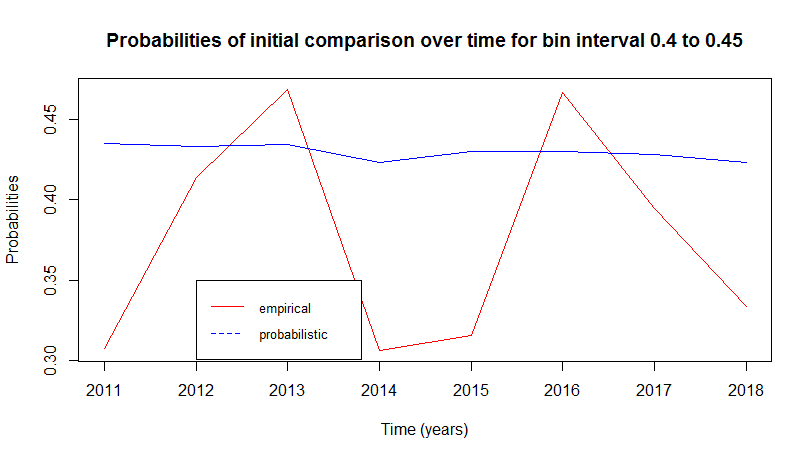


Figure 14: Initial Probabilities for Bin Interval of 0.40 to 0.45 for Pinnacle

It can be seen that, for the bin interval of 0.40 to 0.45 of final odds comparison, probabilities stay constant around the value of 0.44 while the empirical data fluctuates between 0.25 and 0.55, however the empirical line got smoother compared to initial plot which shows that the estimations got better with time. However, still expected probabilities are smoother than real results.

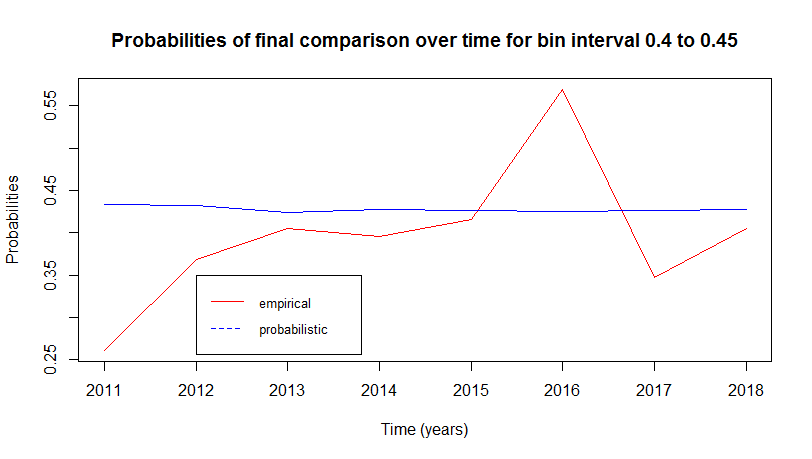


Figure 15: Final Probabilities for Bin Interval of 0.40 to 0.45 for Pinnacle

It can be seen that, for the bin interval of 0.45 to 0.50 of initial odds comparison, probabilities stay constant around the value of 0.48 to 0.46 while the empirical data fluctuates between 0.45 and 0.60. Expected probabilities are smoother than real results. There are empirical values both above and below the probabilistic values, but they are mostly above the probabilistic line and there are a lot of fluctuations.

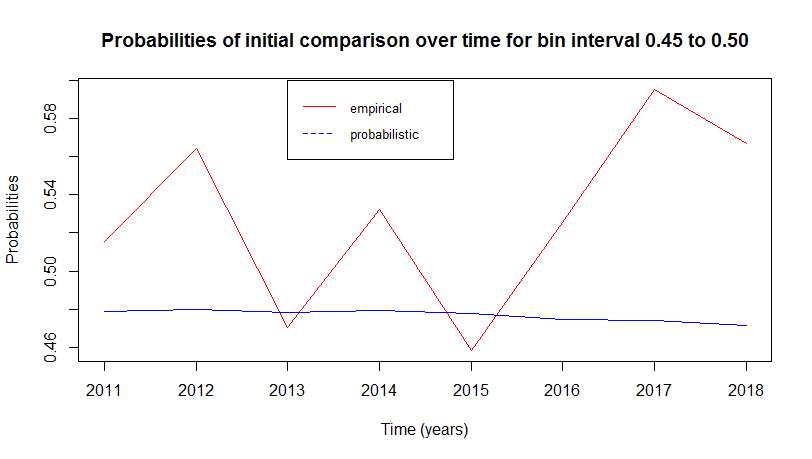


Figure 16: Initial Probabilities for Bin Interval of 0.45 to 0.50 for Pinnacle

It can be seen that, for the bin interval of 0.45 to 0.50 of final odds comparison, probabilities stay constant around the value of 0.47 while the empirical data fluctuates between 0.40 and 0.65. Expected probabilities are smoother than real results. The empirical are still mostly above probabilistic values, but it is smoother, which shows a possible effect of time on probabilities.

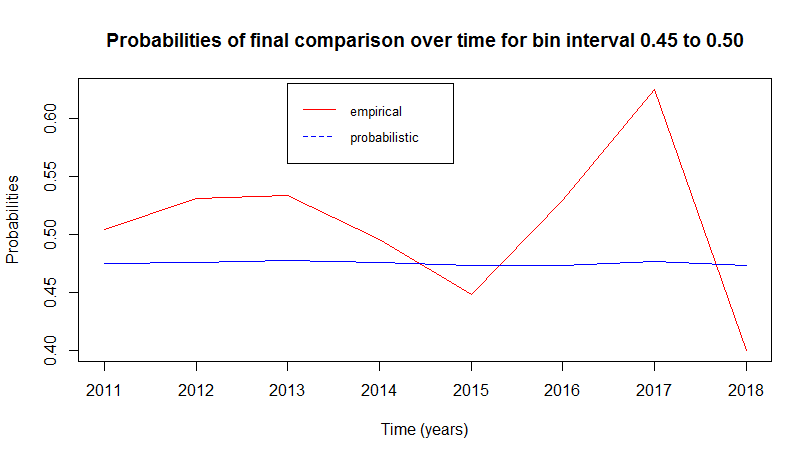


Figure 17: Final Probabilities for Bin Interval of 0.45 to 0.50 for Pinnacle

It can be seen that, for the bin interval of 0.5 to 0.55 of initial odds comparison, probabilities stay constant around the value of 0.50 to 0.52 while the empirical data fluctuates between 0.52 and 0.62. Expected probabilities are smoother than real results. There are empirical values both above and below the probabilistic values, but they are mostly above the probabilistic line.

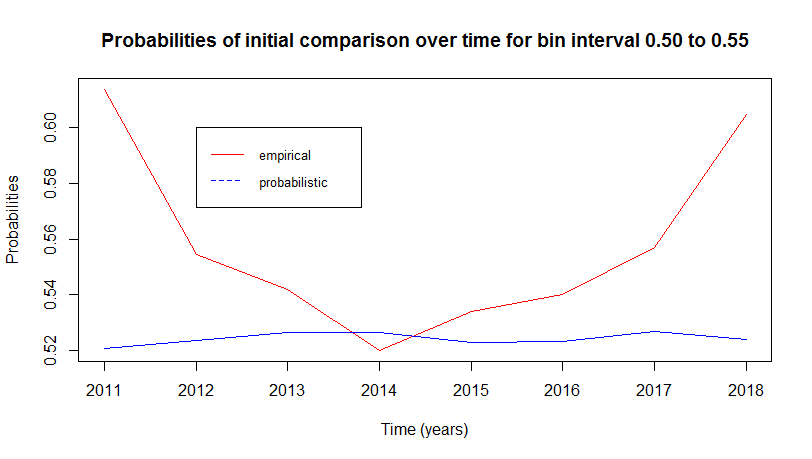


Figure 18: Initial Probabilities for Bin Interval of 0.50 to 0.55 for Pinnacle

It can be seen that, for the bin interval of 0.5 to 0.55 of initial odds comparison, probabilities stay constant around the value of 0.52 while the empirical data fluctuates between 0.45 and 0.62. Expected probabilities are smoother than real results. There are empirical values both above and below the probabilistic values, but they are mostly above the probabilistic line, however the data lower than the probabilistic line increased a little compared to initial time.

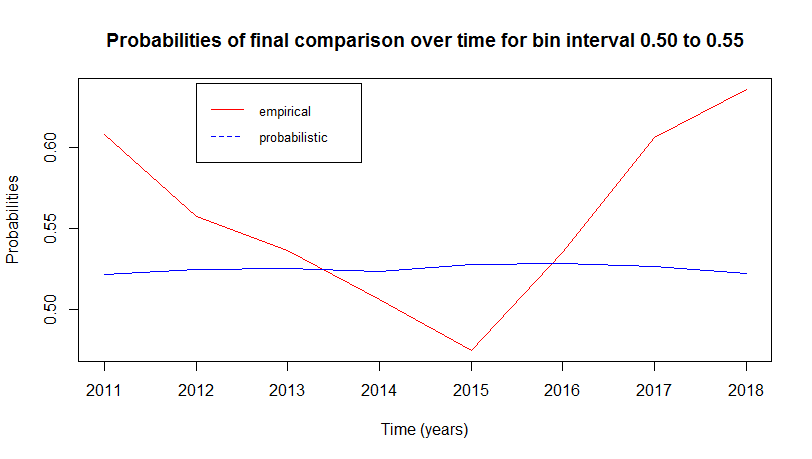


Figure 19: Final Probabilities for Bin Interval of 0.50 to 0.55 for Pinnacle

It can be seen that, for the bin interval of 0.55 to 0.60 of initial odds comparison, probabilities stay constant around the value of 0.57 to 0.60 while the empirical data fluctuates between 0.44 and 0.70. Expected probabilities are smoother than real results. There are empirical values both above and below the probabilistic values, and they are smoother compared to other bin intervals, except the for the year of 2016.

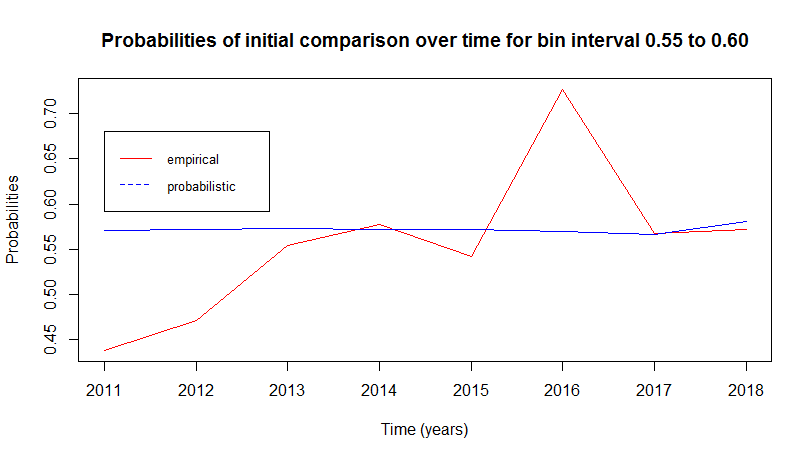


Figure 20: Initial Probabilities for Bin Interval of 0.55 to 0.60 for Pinnacle

It can be seen that, for the bin interval of 0.55 to 0.60 of final odds comparison, probabilities stay constant around the value of 0.57 to 0.60 while the empirical data fluctuates between 0.46 and 0.70. Expected probabilities are smoother than real results. There are empirical values both above and below the probabilistic values, and they are smoother compared to other bin intervals, except the for the year of 2016. The probability for the year of 2014 got away from the empirical data with time.

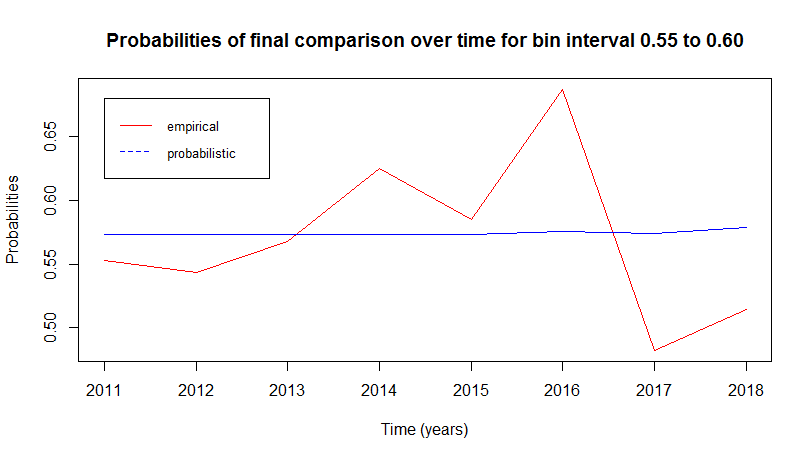


Figure 21: Final Probabilities for Bin Interval of 0.55 to 0.60 for Pinnacle

It can be seen that, for the bin interval of 0.6 to 0.65 of initial odds comparison, probabilities stay constant around the value of 0.64 while the empirical data fluctuates between 0.50 and 0.80. Expected probabilities are smoother than real results. There are empirical values both above and below the probabilistic values and they are more homogeneous compared to previous bin intervals.

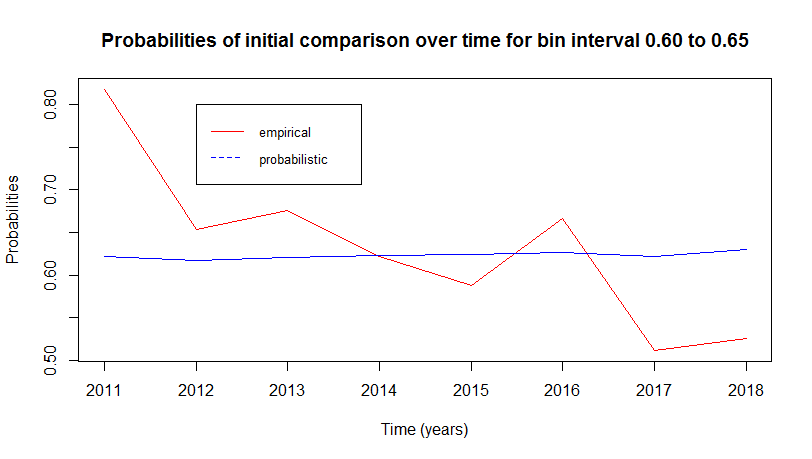


Figure 22: Initial Probabilities for Bin Interval of 0.60 to 0.65 for Pinnacle

It can be seen that, for the bin interval of 0.6 to 0.65 of initial odds comparison, probabilities stay constant around the value of 0.62 while the empirical data fluctuates between 0.56 and 0.70. Expected probabilities are smoother than real results. There are empirical values both above and below the probabilistic values and they are still more homogeneous compared to previous bin intervals even though time changed. The graph seems like empirical line got far away from probabilistic line but it can be seen that it actually got closer when the scale of y axis is checked.

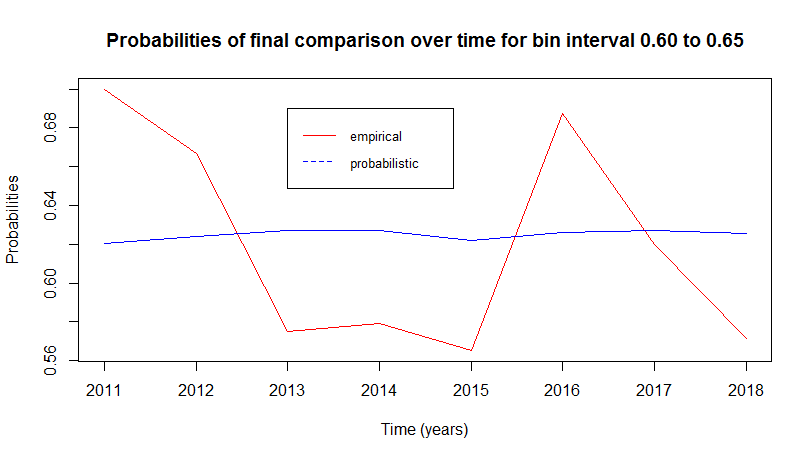


Figure 23: Final Probabilities for Bin Interval of 0.60 to 0.65 for Pinnacle

It can be seen that, for the bin interval of 0.65 to 0.70 of initial odds comparison, probabilities stay constant around the value of 0.65 while the empirical data fluctuates between 0 and 0.10. Expected probabilities are smoother than real results. There are empirical values both above and below the probabilistic values, but it can be seen that this is the least smooth bin among all bins when empirical data is considered.

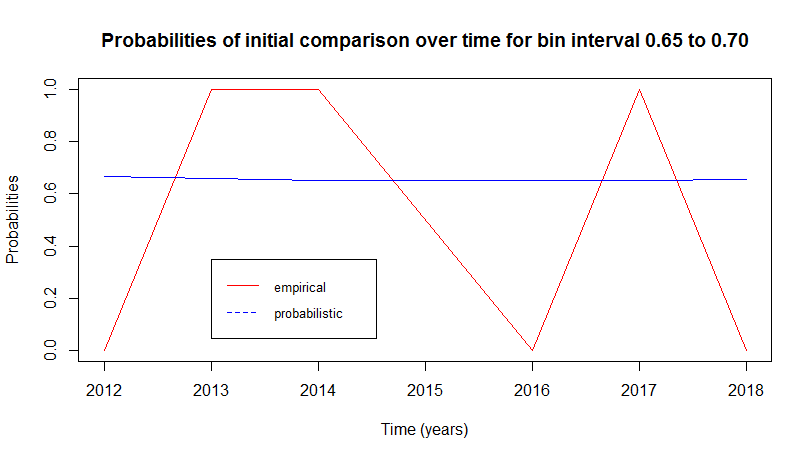


Figure 24: Initial Probabilities for Bin Interval of 0.65 to 0.70 for Pinnacle

It can be seen that, for the bin interval of 0.65 to 0.70 of final odds comparison, probabilities stay constant around the value of 0.65 while the empirical data fluctuates between 0.5 and 1.0. Expected probabilities are smoother than real results. There were empirical values both above and below the probabilistic values initially, and in final time, it can be seen that the empirical values are much smoother compared to initial time, which shows a possible effect of time on probabilities.

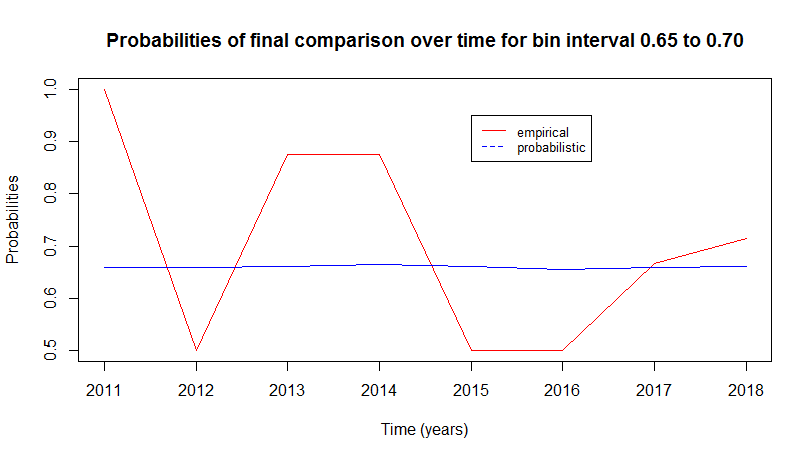


Figure 25: Final Probabilities for Bin Interval of 0.65 to 0.70 for Pinnacle

The pinnacle bookmaker is generally good at deciding over/under bins. And comparison between empirical data line and probabilistic data line over years with time effect on different bins can be seen in the previous graphs.

**Task 2**

In this task changes through time are observed. Initial and final times are considered for comparing over probabilities and 1x2 odds for each year. The code for this part is given in the appendix with all required information and explanations.

First the 1x2 odds are calculated initially and finally and plotted together in the figure below to see their changes over time. It can be seen that some of the matches during the years have higher initial odds compared to their final odds. This can be because of an event that happened to a player with an important role in the game that changed the flow of the match. There are also some matches having higher final odd values compared to initial odds. These are due to the change of possibilities of which team is more likely to win during the match.

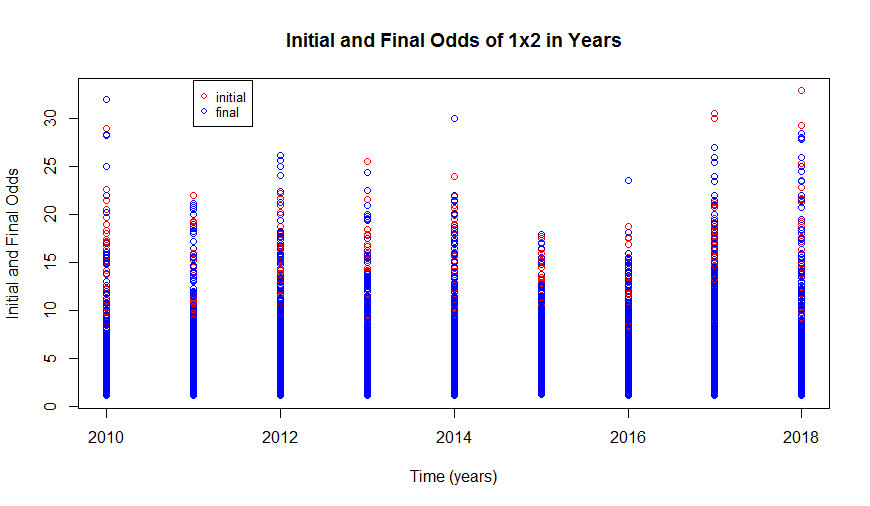


Figure 26: Initial and Final Odds of 1x2 in Years

Another plot is used to compare the initial and final over probabilities for matches in each year. The data calculated in task 1b is used to plot the results and observe their changes over time visually.

The changes in the data points in the figure below is more understandable compared to the previous figure of 1x2 odds. This plot helps us to visualize the results better and see the changes in over probabilities more clearly with time.

From the figure below, it can be seen that some of the matches during the years have higher initial over probabilities compared to their final over probabilities. This can be because of an event that happened to a player with an important role in the game that can change the flow of the match. For example, the bookmaker is sure that home team is going to win. Moreover, the bookmaker guesses that the match result is going to be 3-2. However, an important player of home team gets injured during the match while the score is 1-0. This important player is the one that has the goal record in home team, and without him, they may not be as good. Therefore, the probability of the game finishing with more than 2.5 goals decreases, which causes the over probability to decrease. There are also some matches having lower initial over probabilities compared to their final over probabilities. These are due to the increase of the possibility of having more goals during the match. The reason may be a player being better than expected that day and having more motivation to win.

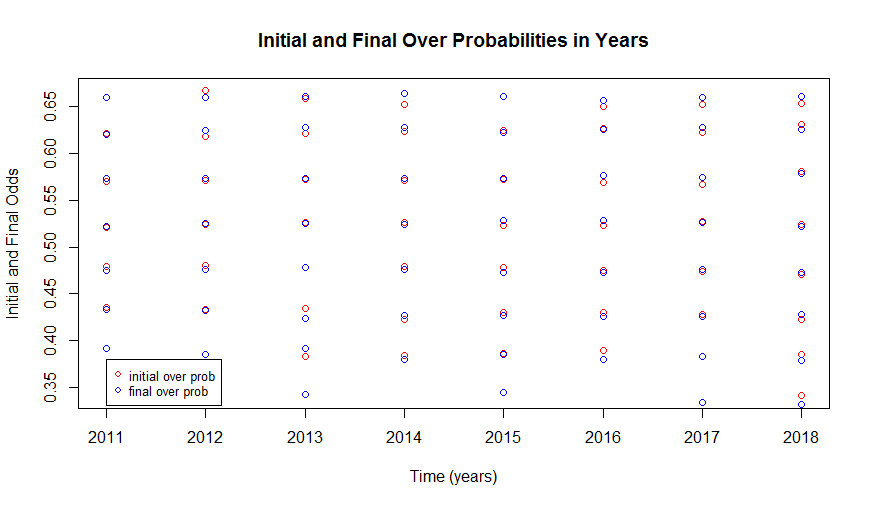


Figure 27: Initial and Final Over Probabilities in Years

There is no data points under the legend box in the figure above, so all information is visible, there is no loss of data observation.

**Appendix**

#gizem kurtbay- code for hw 1 of ie582

#install.packages("data.table")

#install.packages("anytime")

require(data.table)

require(anytime)

matches\_file\_path='C:/Users/Gizem/Downloads/ie582icin/df9b1196-e3cf-4cc7-9159-f236fe738215\_matches.rds'

odd\_details\_file\_path='C:/Users/Gizem/Downloads/ie582icin/df9b1196-e3cf-4cc7-9159-f236fe738215\_odd\_details.rds'

matches=readRDS(matches\_file\_path)

odds=readRDS(odd\_details\_file\_path)

str(matches)

matches=unique(matches)

zero\_zero\_matches=matches[score=='0:0']

#transform unix time to date

matches[,match\_date:=anydate(date)]

#transform unix time to date time

matches[,match\_time:=anytime(date)]

#order by home team and match date (decreasing)

matches=matches[order(home,-match\_time)]

#only one column deletion

matches[,date:=NULL]

#only delete any number of columns

matches[,c("match\_date","date"):=NULL]

#matches[,3,with=F]

matches[,c("HomeGoals","AwayGoals"):=tstrsplit(score,':')]

matches[,Year:=year(match\_time)]

matches[,Month:=month(match\_time)]

matches[,Weekday:=wday(match\_time)]

matches[,Hour:=hour(match\_time)]

#transform characters to numeric for scores

matches$HomeGoals=as.numeric(matches$HomeGoals)

matches[,AwayGoals:=as.numeric(AwayGoals)]

#calculate total goals

matches[,TotalGoals:=HomeGoals+AwayGoals]

# mark over under

matches[,IsOver:=0]

matches[TotalGoals>2,IsOver:=1]

#alternative

matches[,IsOverAlt:=as.numeric(TotalGoals>2)]

#filter na scores

matches[is.na(score)]

#filter all NAs

matches=matches[complete.cases(matches)]

#yearly average goals

yearly\_goals=matches[,list(AvgGoals=mean(TotalGoals)),by=list(Year)]

#yearly hourly average goals

yearly\_hourly\_goals=matches[,list(AvgGoals=mean(TotalGoals),

MaxGoals=max(TotalGoals),

NGames=.N, AltNGames=length(matchId)),

by=list(Year,Hour)]

#get the game with the max total goals

max\_game=matches[,list(MaxGoals=max(TotalGoals),

homeMax=home[which.max(TotalGoals)],

awayMax=away[which.max(TotalGoals)]),

by=list(Year,Hour)]

#get the game with the max total goals for Year greater than 2017

max\_game=matches[Year>2017,list(MaxGoals=max(TotalGoals),

homeMax=home[which.max(TotalGoals)],

awayMax=away[which.max(TotalGoals)]),

by=list(Year,Hour)]

#filter over under 2.5

odds\_ov\_un=odds[betType=='ou' & totalhandicap=='2.5']

#remove total handicap

odds\_ov\_un[,totalhandicap:=NULL]

#remove original odds data

#rm(odds); gc();

#order data in ascending date

odds\_ov\_un=odds\_ov\_un[order(matchId, oddtype,bookmaker,date)]

odds\_ov\_un\_initial=odds\_ov\_un[,list(start\_odd=odd[1]),

by=list(matchId,oddtype,bookmaker)]

odds\_ov\_un\_final=odds\_ov\_un[,list(final\_odd=odd[.N]),

by=list(matchId,oddtype,bookmaker)]

#transform to wide format

wide\_odds\_initial=dcast(odds\_ov\_un\_initial,

matchId~oddtype+bookmaker,

value.var='start\_odd')

#transform to long

long\_ov\_un\_initial=melt(wide\_odds\_initial,id.vars=1,

measure.vars=2:ncol(wide\_odds\_initial))

## FOR 5 BOOKMAKERS!!!

##for pinnacle

#get pinnacle initial over under odds

pinnacle\_over\_under\_initial=odds\_ov\_un\_initial[bookmaker=='Pinnacle']

pinnacle\_wide\_initial=dcast (pinnacle\_over\_under\_initial,

matchId~oddtype,

value.var='start\_odd')

# join odds with matches for pinnacle

merged\_matches\_pinnacle\_initial=merge(matches,pinnacle\_wide\_initial,by='matchId' )

merged\_matches\_pinnacle\_initial[,probOver:=1/over]

merged\_matches\_pinnacle\_initial[,probUnder:=1/under]

merged\_matches\_pinnacle\_initial[,totalProb:=probOver+probUnder]

merged\_matches\_pinnacle\_initial[,probOver:=probOver/totalProb]

merged\_matches\_pinnacle\_initial[,probUnder:=probUnder/totalProb]

merged\_matches\_pinnacle\_initial=merged\_matches\_pinnacle\_initial[complete.cases(merged\_matches\_pinnacle\_initial)]

merged\_matches\_pinnacle\_initial[,totalProb:=NULL]

cutpoints=seq(0,1,0.05)

merged\_matches\_pinnacle\_initial[,odd\_cut\_over:=cut(probOver,cutpoints)]

summary\_table\_pinnacle\_initial=merged\_matches\_pinnacle\_initial[,list(empirical\_over\_pinnacle=mean(IsOver),

probabilistic\_over\_pinnacle=mean(probOver),.N),

by=list(Year,odd\_cut\_over)]

summary\_table\_pinnacle\_initial=summary\_table\_pinnacle\_initial[order(Year)]

plot(summary\_table\_pinnacle\_initial[,list(empirical\_over\_pinnacle,probabilistic\_over\_pinnacle)],cex=1,main='pinnacle initial odds')

abline(0,1,col='red')

#get pinnacle final over under odds

pinnacle\_over\_under\_final=odds\_ov\_un\_final[bookmaker=='Pinnacle']

pinnacle\_wide\_final=dcast (pinnacle\_over\_under\_final,

matchId~oddtype,

value.var='final\_odd')

# join odds with matches for pinnacle

merged\_matches\_pinnacle\_final=merge(matches,pinnacle\_wide\_final,by='matchId' )

merged\_matches\_pinnacle\_final[,probOver:=1/over]

merged\_matches\_pinnacle\_final[,probUnder:=1/under]

merged\_matches\_pinnacle\_final[,totalProb:=probOver+probUnder]

merged\_matches\_pinnacle\_final[,probOver:=probOver/totalProb]

merged\_matches\_pinnacle\_final[,probUnder:=probUnder/totalProb]

merged\_matches\_pinnacle\_final=merged\_matches\_pinnacle\_final[complete.cases(merged\_matches\_pinnacle\_final)]

merged\_matches\_pinnacle\_final[,totalProb:=NULL]

cutpoints=seq(0,1,0.05)

merged\_matches\_pinnacle\_final[,odd\_cut\_over:=cut(probOver,cutpoints)]

summary\_table\_pinnacle\_final=merged\_matches\_pinnacle\_final[,list(empirical\_over\_pinnacle=mean(IsOver),

probabilistic\_over\_pinnacle=mean(probOver),.N),

by=list(Year,odd\_cut\_over)]

summary\_table\_pinnacle\_final=summary\_table\_pinnacle\_final[order(Year)]

plot(summary\_table\_pinnacle\_final[,list(empirical\_over\_pinnacle,probabilistic\_over\_pinnacle)],cex=1,main='pinnacle final odds')

abline(0,1,col='red')

##for bet365

#get bet365 initial over under odds

bet365\_over\_under\_initial=odds\_ov\_un\_initial[bookmaker=='bet365']

bet365\_wide\_initial=dcast (bet365\_over\_under\_initial,

matchId~oddtype,

value.var='start\_odd')

# join odds with matches for bet365

merged\_matches\_bet365\_initial=merge(matches,bet365\_wide\_initial,by='matchId' )

merged\_matches\_bet365\_initial[,probOver:=1/over]

merged\_matches\_bet365\_initial[,probUnder:=1/under]

merged\_matches\_bet365\_initial[,totalProb:=probOver+probUnder]

merged\_matches\_bet365\_initial[,probOver:=probOver/totalProb]

merged\_matches\_bet365\_initial[,probUnder:=probUnder/totalProb]

merged\_matches\_bet365\_initial=merged\_matches\_bet365\_initial[complete.cases(merged\_matches\_bet365\_initial)]

merged\_matches\_bet365\_initial[,totalProb:=NULL]

cutpoints=seq(0,1,0.05)

merged\_matches\_bet365\_initial[,odd\_cut\_over:=cut(probOver,cutpoints)]

summary\_table\_bet365\_initial=merged\_matches\_bet365\_initial[,list(empirical\_over\_bet365=mean(IsOver),

probabilistic\_over\_bet365=mean(probOver),.N),

by=list(Year,odd\_cut\_over)]

summary\_table\_bet365\_initial=summary\_table\_bet365\_initial[order(Year)]

plot(summary\_table\_bet365\_initial[,list(empirical\_over\_bet365,probabilistic\_over\_bet365)],cex=1,main='bet365 initial odds')

abline(0,1,col='red')

#get bet365 final over under odds

bet365\_over\_under\_final=odds\_ov\_un\_final[bookmaker=='bet365']

bet365\_wide\_final=dcast (bet365\_over\_under\_final,

matchId~oddtype,

value.var='final\_odd')

# join odds with matches for bet365

merged\_matches\_bet365\_final=merge(matches,bet365\_wide\_final,by='matchId' )

merged\_matches\_bet365\_final[,probOver:=1/over]

merged\_matches\_bet365\_final[,probUnder:=1/under]

merged\_matches\_bet365\_final[,totalProb:=probOver+probUnder]

merged\_matches\_bet365\_final[,probOver:=probOver/totalProb]

merged\_matches\_bet365\_final[,probUnder:=probUnder/totalProb]

merged\_matches\_bet365\_final=merged\_matches\_bet365\_final[complete.cases(merged\_matches\_bet365\_final)]

merged\_matches\_bet365\_final[,totalProb:=NULL]

cutpoints=seq(0,1,0.05)

merged\_matches\_bet365\_final[,odd\_cut\_over:=cut(probOver,cutpoints)]

summary\_table\_bet365\_final=merged\_matches\_bet365\_final[,list(empirical\_over\_bet365=mean(IsOver),

probabilistic\_over\_bet365=mean(probOver),.N),

by=list(Year,odd\_cut\_over)]

summary\_table\_bet365\_final=summary\_table\_bet365\_final[order(Year)]

plot(summary\_table\_bet365\_final[,list(empirical\_over\_bet365,probabilistic\_over\_bet365)],cex=1,main='bet365 final odds')

abline(0,1,col='red')

##for Betway

#get Betway initial over under odds

Betway\_over\_under\_initial=odds\_ov\_un\_initial[bookmaker=='Betway']

Betway\_wide\_initial=dcast (Betway\_over\_under\_initial,

matchId~oddtype,

value.var='start\_odd')

# join odds with matches for Betway

merged\_matches\_Betway\_initial=merge(matches,Betway\_wide\_initial,by='matchId' )

merged\_matches\_Betway\_initial[,probOver:=1/over]

merged\_matches\_Betway\_initial[,probUnder:=1/under]

merged\_matches\_Betway\_initial[,totalProb:=probOver+probUnder]

merged\_matches\_Betway\_initial[,probOver:=probOver/totalProb]

merged\_matches\_Betway\_initial[,probUnder:=probUnder/totalProb]

merged\_matches\_Betway\_initial=merged\_matches\_Betway\_initial[complete.cases(merged\_matches\_Betway\_initial)]

merged\_matches\_Betway\_initial[,totalProb:=NULL]

cutpoints=seq(0,1,0.05)

merged\_matches\_Betway\_initial[,odd\_cut\_over:=cut(probOver,cutpoints)]

summary\_table\_Betway\_initial=merged\_matches\_Betway\_initial[,list(empirical\_over\_Betway=mean(IsOver),

probabilistic\_over\_Betway=mean(probOver),.N),

by=list(Year,odd\_cut\_over)]

summary\_table\_Betway\_initial=summary\_table\_Betway\_initial[order(Year)]

plot(summary\_table\_Betway\_initial[,list(empirical\_over\_Betway,probabilistic\_over\_Betway)],cex=1,main='Betway initial odds')

abline(0,1,col='red')

#get Betway final over under odds

Betway\_over\_under\_final=odds\_ov\_un\_final[bookmaker=='Betway']

Betway\_wide\_final=dcast (Betway\_over\_under\_final,

matchId~oddtype,

value.var='final\_odd')

# join odds with matches for Betway

merged\_matches\_Betway\_final=merge(matches,Betway\_wide\_final,by='matchId' )

merged\_matches\_Betway\_final[,probOver:=1/over]

merged\_matches\_Betway\_final[,probUnder:=1/under]

merged\_matches\_Betway\_final[,totalProb:=probOver+probUnder]

merged\_matches\_Betway\_final[,probOver:=probOver/totalProb]

merged\_matches\_Betway\_final[,probUnder:=probUnder/totalProb]

merged\_matches\_Betway\_final=merged\_matches\_Betway\_final[complete.cases(merged\_matches\_Betway\_final)]

merged\_matches\_Betway\_final[,totalProb:=NULL]

cutpoints=seq(0,1,0.05)

merged\_matches\_Betway\_final[,odd\_cut\_over:=cut(probOver,cutpoints)]

summary\_table\_Betway\_final=merged\_matches\_Betway\_final[,list(empirical\_over\_Betway=mean(IsOver),

probabilistic\_over\_Betway=mean(probOver),.N),

by=list(Year,odd\_cut\_over)]

summary\_table\_Betway\_final=summary\_table\_Betway\_final[order(Year)]

plot(summary\_table\_Betway\_final[,list(empirical\_over\_Betway,probabilistic\_over\_Betway)],cex=1,main='Betway final odds')

abline(0,1,col='red')

##for bwin

#get bwin initial over under odds

bwin\_over\_under\_initial=odds\_ov\_un\_initial[bookmaker=='bwin']

bwin\_wide\_initial=dcast (bwin\_over\_under\_initial,

matchId~oddtype,

value.var='start\_odd')

# join odds with matches for bwin

merged\_matches\_bwin\_initial=merge(matches,bwin\_wide\_initial,by='matchId' )

merged\_matches\_bwin\_initial[,probOver:=1/over]

merged\_matches\_bwin\_initial[,probUnder:=1/under]

merged\_matches\_bwin\_initial[,totalProb:=probOver+probUnder]

merged\_matches\_bwin\_initial[,probOver:=probOver/totalProb]

merged\_matches\_bwin\_initial[,probUnder:=probUnder/totalProb]

merged\_matches\_bwin\_initial=merged\_matches\_bwin\_initial[complete.cases(merged\_matches\_bwin\_initial)]

merged\_matches\_bwin\_initial[,totalProb:=NULL]

cutpoints=seq(0,1,0.05)

merged\_matches\_bwin\_initial[,odd\_cut\_over:=cut(probOver,cutpoints)]

summary\_table\_bwin\_initial=merged\_matches\_bwin\_initial[,list(empirical\_over\_bwin=mean(IsOver),

probabilistic\_over\_bwin=mean(probOver),.N),

by=list(Year,odd\_cut\_over)]

summary\_table\_bwin\_initial=summary\_table\_bwin\_initial[order(Year)]

plot(summary\_table\_bwin\_initial[,list(empirical\_over\_bwin,probabilistic\_over\_bwin)],cex=1,main='bwin initial odds')

abline(0,1,col='red')

#get bwin final over under odds

bwin\_over\_under\_final=odds\_ov\_un\_final[bookmaker=='bwin']

bwin\_wide\_final=dcast (bwin\_over\_under\_final,

matchId~oddtype,

value.var='final\_odd')

# join odds with matches for bwin

merged\_matches\_bwin\_final=merge(matches,bwin\_wide\_final,by='matchId' )

merged\_matches\_bwin\_final[,probOver:=1/over]

merged\_matches\_bwin\_final[,probUnder:=1/under]

merged\_matches\_bwin\_final[,totalProb:=probOver+probUnder]

merged\_matches\_bwin\_final[,probOver:=probOver/totalProb]

merged\_matches\_bwin\_final[,probUnder:=probUnder/totalProb]

merged\_matches\_bwin\_final=merged\_matches\_bwin\_final[complete.cases(merged\_matches\_bwin\_final)]

merged\_matches\_bwin\_final[,totalProb:=NULL]

cutpoints=seq(0,1,0.05)

merged\_matches\_bwin\_final[,odd\_cut\_over:=cut(probOver,cutpoints)]

summary\_table\_bwin\_final=merged\_matches\_bwin\_final[,list(empirical\_over\_bwin=mean(IsOver),

probabilistic\_over\_bwin=mean(probOver),.N),

by=list(Year,odd\_cut\_over)]

summary\_table\_bwin\_final=summary\_table\_bwin\_final[order(Year)]

plot(summary\_table\_bwin\_final[,list(empirical\_over\_bwin,probabilistic\_over\_bwin)],cex=1,main='bwin final odds')

abline(0,1,col='red')

##for Betsson

#get Betsson initial over under odds

Betsson\_over\_under\_initial=odds\_ov\_un\_initial[bookmaker=='Betsson']

Betsson\_wide\_initial=dcast (Betsson\_over\_under\_initial,

matchId~oddtype,

value.var='start\_odd')

# join odds with matches for Betsson

merged\_matches\_Betsson\_initial=merge(matches,Betsson\_wide\_initial,by='matchId' )

merged\_matches\_Betsson\_initial[,probOver:=1/over]

merged\_matches\_Betsson\_initial[,probUnder:=1/under]

merged\_matches\_Betsson\_initial[,totalProb:=probOver+probUnder]

merged\_matches\_Betsson\_initial[,probOver:=probOver/totalProb]

merged\_matches\_Betsson\_initial[,probUnder:=probUnder/totalProb]

merged\_matches\_Betsson\_initial=merged\_matches\_Betsson\_initial[complete.cases(merged\_matches\_Betsson\_initial)]

merged\_matches\_Betsson\_initial[,totalProb:=NULL]

cutpoints=seq(0,1,0.05)

merged\_matches\_Betsson\_initial[,odd\_cut\_over:=cut(probOver,cutpoints)]

summary\_table\_Betsson\_initial=merged\_matches\_Betsson\_initial[,list(empirical\_over\_Betsson=mean(IsOver),

probabilistic\_over\_Betsson=mean(probOver),.N),

by=list(Year,odd\_cut\_over)]

summary\_table\_Betsson\_initial=summary\_table\_Betsson\_initial[order(Year)]

plot(summary\_table\_Betsson\_initial[,list(empirical\_over\_Betsson,probabilistic\_over\_Betsson)],cex=1,main='Betsson initial odds')

abline(0,1,col='red')

#get Betsson final over under odds

Betsson\_over\_under\_final=odds\_ov\_un\_final[bookmaker=='Betsson']

Betsson\_wide\_final=dcast (Betsson\_over\_under\_final,

matchId~oddtype,

value.var='final\_odd')

# join odds with matches for Betsson

merged\_matches\_Betsson\_final=merge(matches,Betsson\_wide\_final,by='matchId' )

merged\_matches\_Betsson\_final[,probOver:=1/over]

merged\_matches\_Betsson\_final[,probUnder:=1/under]

merged\_matches\_Betsson\_final[,totalProb:=probOver+probUnder]

merged\_matches\_Betsson\_final[,probOver:=probOver/totalProb]

merged\_matches\_Betsson\_final[,probUnder:=probUnder/totalProb]

merged\_matches\_Betsson\_final=merged\_matches\_Betsson\_final[complete.cases(merged\_matches\_Betsson\_final)]

merged\_matches\_Betsson\_final[,totalProb:=NULL]

cutpoints=seq(0,1,0.05)

merged\_matches\_Betsson\_final[,odd\_cut\_over:=cut(probOver,cutpoints)]

summary\_table\_Betsson\_final=merged\_matches\_Betsson\_final[,list(empirical\_over\_Betsson=mean(IsOver),

probabilistic\_over\_Betsson=mean(probOver),.N),

by=list(Year,odd\_cut\_over)]

summary\_table\_Betsson\_final=summary\_table\_Betsson\_final[order(Year)]

plot(summary\_table\_Betsson\_final[,list(empirical\_over\_Betsson,probabilistic\_over\_Betsson)],cex=1,main='Betsson final odds')

abline(0,1,col='red')

# Task 1 b

#Pinnacle is selected as the bookmaker for task1b

#bin = 0.35 to 0.40 - initial

merged\_matches\_35\_40\_initial=merged\_matches\_pinnacle\_initial[odd\_cut\_over=="(0.35,0.4]"]

merged\_matches\_35\_40\_initial=merged\_matches\_35\_40\_initial[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_35\_40\_initial=merged\_matches\_35\_40\_initial[order(Year)]

plot(merged\_matches\_35\_40\_initial$Year,merged\_matches\_35\_40\_initial$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of initial comparison over time for bin interval 0.35 to 0.40')

lines(merged\_matches\_35\_40\_initial$Year,merged\_matches\_35\_40\_initial$probabilistic\_over,col="blue")

legend(2014,0.20, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.4 to 0.45 - initial

merged\_matches\_40\_45\_initial=merged\_matches\_pinnacle\_initial[odd\_cut\_over=="(0.4,0.45]"]

merged\_matches\_40\_45\_initial=merged\_matches\_40\_45\_initial[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_40\_45\_initial=merged\_matches\_40\_45\_initial[order(Year)]

plot(merged\_matches\_40\_45\_initial$Year,merged\_matches\_40\_45\_initial$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of initial comparison over time for bin interval 0.4 to 0.45')

lines(merged\_matches\_40\_45\_initial$Year,merged\_matches\_40\_45\_initial$probabilistic\_over,col="blue")

legend(2012,0.35, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.45 to 0.50 - initial

merged\_matches\_45\_50\_initial=merged\_matches\_pinnacle\_initial[odd\_cut\_over=="(0.45,0.5]"]

merged\_matches\_45\_50\_initial=merged\_matches\_45\_50\_initial[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_45\_50\_initial=merged\_matches\_45\_50\_initial[order(Year)]

plot(merged\_matches\_45\_50\_initial$Year,merged\_matches\_45\_50\_initial$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of initial comparison over time for bin interval 0.45 to 0.50')

lines(merged\_matches\_45\_50\_initial$Year,merged\_matches\_45\_50\_initial$probabilistic\_over,col="blue")

legend(2013,0.60, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.50 to 0.55 - initial

merged\_matches\_50\_55\_initial=merged\_matches\_pinnacle\_initial[odd\_cut\_over=="(0.5,0.55]"]

merged\_matches\_50\_55\_initial=merged\_matches\_50\_55\_initial[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_50\_55\_initial=merged\_matches\_50\_55\_initial[order(Year)]

plot(merged\_matches\_50\_55\_initial$Year,merged\_matches\_50\_55\_initial$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of initial comparison over time for bin interval 0.50 to 0.55')

lines(merged\_matches\_50\_55\_initial$Year,merged\_matches\_50\_55\_initial$probabilistic\_over,col="blue")

legend(2012,0.60, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.55 to 0.60 - initial

merged\_matches\_55\_60\_initial=merged\_matches\_pinnacle\_initial[odd\_cut\_over=="(0.55,0.6]"]

merged\_matches\_55\_60\_initial=merged\_matches\_55\_60\_initial[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_55\_60\_initial=merged\_matches\_55\_60\_initial[order(Year)]

plot(merged\_matches\_55\_60\_initial$Year,merged\_matches\_55\_60\_initial$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of initial comparison over time for bin interval 0.55 to 0.60')

lines(merged\_matches\_55\_60\_initial$Year,merged\_matches\_55\_60\_initial$probabilistic\_over,col="blue")

legend(2011,0.68, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.60 to 0.65 - initial

merged\_matches\_60\_65\_initial=merged\_matches\_pinnacle\_initial[odd\_cut\_over=="(0.6,0.65]"]

merged\_matches\_60\_65\_initial=merged\_matches\_60\_65\_initial[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_60\_65\_initial=merged\_matches\_60\_65\_initial[order(Year)]

plot(merged\_matches\_60\_65\_initial$Year,merged\_matches\_60\_65\_initial$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of initial comparison over time for bin interval 0.60 to 0.65')

lines(merged\_matches\_60\_65\_initial$Year,merged\_matches\_60\_65\_initial$probabilistic\_over,col="blue")

legend(2012,0.80, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.65 to 0.70 - initial

merged\_matches\_65\_70\_initial=merged\_matches\_pinnacle\_initial[odd\_cut\_over=="(0.65,0.7]"]

merged\_matches\_65\_70\_initial=merged\_matches\_65\_70\_initial[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_65\_70\_initial=merged\_matches\_65\_70\_initial[order(Year)]

plot(merged\_matches\_65\_70\_initial$Year,merged\_matches\_65\_70\_initial$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of initial comparison over time for bin interval 0.65 to 0.70')

lines(merged\_matches\_65\_70\_initial$Year,merged\_matches\_65\_70\_initial$probabilistic\_over,col="blue")

legend(2013,0.35, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.30 to 0.35 - final

merged\_matches\_30\_35\_final=merged\_matches\_pinnacle\_final[odd\_cut\_over=="(0.3,0.35]"]

merged\_matches\_30\_35\_final=merged\_matches\_30\_35\_final[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_30\_35\_final=merged\_matches\_30\_35\_final[order(Year)]

plot(merged\_matches\_30\_35\_final$Year,merged\_matches\_30\_35\_final$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of final comparison over time for bin interval 0.30 to 0.35')

lines(merged\_matches\_30\_35\_final$Year,merged\_matches\_30\_35\_final$probabilistic\_over,col="blue")

legend(2015,0.10, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.35 to 0.40 - final

merged\_matches\_35\_40\_final=merged\_matches\_pinnacle\_final[odd\_cut\_over=="(0.35,0.4]"]

merged\_matches\_35\_40\_final=merged\_matches\_35\_40\_final[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_35\_40\_final=merged\_matches\_35\_40\_final[order(Year)]

plot(merged\_matches\_35\_40\_final$Year,merged\_matches\_35\_40\_final$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of final comparison over time for bin interval 0.35 to 0.40')

lines(merged\_matches\_35\_40\_final$Year,merged\_matches\_35\_40\_final$probabilistic\_over,col="blue")

legend(2014,0.20, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.4 to 0.45 - final

merged\_matches\_40\_45\_final=merged\_matches\_pinnacle\_final[odd\_cut\_over=="(0.4,0.45]"]

merged\_matches\_40\_45\_final=merged\_matches\_40\_45\_final[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_40\_45\_final=merged\_matches\_40\_45\_final[order(Year)]

plot(merged\_matches\_40\_45\_final$Year,merged\_matches\_40\_45\_final$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of final comparison over time for bin interval 0.4 to 0.45')

lines(merged\_matches\_40\_45\_final$Year,merged\_matches\_40\_45\_final$probabilistic\_over,col="blue")

legend(2012,0.35, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.45 to 0.50 - final

merged\_matches\_45\_50\_final=merged\_matches\_pinnacle\_final[odd\_cut\_over=="(0.45,0.5]"]

merged\_matches\_45\_50\_final=merged\_matches\_45\_50\_final[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_45\_50\_final=merged\_matches\_45\_50\_final[order(Year)]

plot(merged\_matches\_45\_50\_final$Year,merged\_matches\_45\_50\_final$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of final comparison over time for bin interval 0.45 to 0.50')

lines(merged\_matches\_45\_50\_final$Year,merged\_matches\_45\_50\_final$probabilistic\_over,col="blue")

legend(2013,0.63, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.50 to 0.55 - final

merged\_matches\_50\_55\_final=merged\_matches\_pinnacle\_final[odd\_cut\_over=="(0.5,0.55]"]

merged\_matches\_50\_55\_final=merged\_matches\_50\_55\_final[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_50\_55\_final=merged\_matches\_50\_55\_final[order(Year)]

plot(merged\_matches\_50\_55\_final$Year,merged\_matches\_50\_55\_final$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of final comparison over time for bin interval 0.50 to 0.55')

lines(merged\_matches\_50\_55\_final$Year,merged\_matches\_50\_55\_final$probabilistic\_over,col="blue")

legend(2012,0.64, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.55 to 0.60 - final

merged\_matches\_55\_60\_final=merged\_matches\_pinnacle\_final[odd\_cut\_over=="(0.55,0.6]"]

merged\_matches\_55\_60\_final=merged\_matches\_55\_60\_final[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_55\_60\_final=merged\_matches\_55\_60\_final[order(Year)]

plot(merged\_matches\_55\_60\_final$Year,merged\_matches\_55\_60\_final$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of final comparison over time for bin interval 0.55 to 0.60')

lines(merged\_matches\_55\_60\_final$Year,merged\_matches\_55\_60\_final$probabilistic\_over,col="blue")

legend(2011,0.68, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.60 to 0.65 - final

merged\_matches\_60\_65\_final=merged\_matches\_pinnacle\_final[odd\_cut\_over=="(0.6,0.65]"]

merged\_matches\_60\_65\_final=merged\_matches\_60\_65\_final[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_60\_65\_final=merged\_matches\_60\_65\_final[order(Year)]

plot(merged\_matches\_60\_65\_final$Year,merged\_matches\_60\_65\_final$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of final comparison over time for bin interval 0.60 to 0.65')

lines(merged\_matches\_60\_65\_final$Year,merged\_matches\_60\_65\_final$probabilistic\_over,col="blue")

legend(2013,0.69, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

#bin = 0.65 to 0.70 - final

merged\_matches\_65\_70\_final=merged\_matches\_pinnacle\_final[odd\_cut\_over=="(0.65,0.7]"]

merged\_matches\_65\_70\_final=merged\_matches\_65\_70\_final[,list(empirical\_over=mean(IsOver),

probabilistic\_over=mean(probOver),.N),by=list(Year)]

merged\_matches\_65\_70\_final=merged\_matches\_65\_70\_final[order(Year)]

plot(merged\_matches\_65\_70\_final$Year,merged\_matches\_65\_70\_final$empirical\_over,type = "l",col="red",xlab="Time (years)",ylab="Probabilities"

,main='Probabilities of final comparison over time for bin interval 0.65 to 0.70')

lines(merged\_matches\_65\_70\_final$Year,merged\_matches\_65\_70\_final$probabilistic\_over,col="blue")

legend(2015,0.95, legend=c("empirical", "probabilistic"), col=c("red", "blue"), lty=1:2, cex=0.8)

# Task 2

#Pinnacle was selected

pinnacle\_odds\_1x2=odds[betType=='1x2'& bookmaker=="Pinnacle"]

pinnacle\_odds\_1x2\_initial=pinnacle\_odds\_1x2[,list(start\_odd=odd[1]),

by=list(matchId,oddtype,bookmaker)]

pinnacle\_odds\_1x2\_final=pinnacle\_odds\_1x2[,list(final\_odd=odd[.N]),

by=list(matchId,oddtype,bookmaker)]

#total odd is created from initial and final odds by merge

pinnacle\_odds\_1x2\_total<-merge(pinnacle\_odds\_1x2\_initial,pinnacle\_odds\_1x2\_final,by=.EACHI)

pinnacle\_odds\_1x2\_total<-merge(pinnacle\_odds\_1x2\_total,matches,by="matchId")

#plots are coded for visualization of results

#Initial and Final Odds of 1x2 are plotted for Years

plot(pinnacle\_odds\_1x2\_total$Year,pinnacle\_odds\_1x2\_total$start\_odd,type = "p",col="red",xlab="Time (years)",ylab="Initial and Final Odds"

,main='Initial and Final Odds of 1x2 in Years')

points(pinnacle\_odds\_1x2\_total$Year,pinnacle\_odds\_1x2\_total$final\_odd,col="blue")

legend(2011,34, legend=c("initial", "final"), col=c("red", "blue"),pch = c(1, 1), cex=0.8)

#Initial and Final probabilities of over are plotted for Years

plot(summary\_table\_pinnacle\_initial$Year,summary\_table\_pinnacle\_initial$probabilistic\_over\_pinnacle,type = "p",col="red",xlab="Time (years)",ylab="Initial and Final Odds"

,main='Initial and Final Over Probabilities in Years')

points(summary\_table\_pinnacle\_final$Year,summary\_table\_pinnacle\_final$probabilistic\_over\_pinnacle,col="blue")

legend(2011,0.38, legend=c("initial over prob", "final over prob"), col=c("red", "blue"),pch = c(1, 1) , cex=0.8)

#there is no data points under the legend box in this plot, so all information is visible.