#####################################################################

#Yuhao Wang 03/27/2018 data cleanning process KUZ7

#####################################################################

#import data, necessary functions, and library

#function index0: gives new index for matrix

#function slincing: take only price out of raw data

#####################################################################

library("highfrequency")

KUZ = read.csv("C:/Users/dell/Desktop/821project/KUZ7\_wyh.csv")

index0 <- function(x){

rownames(x) <- 1:dim(x)[1]

return(x)

}

#####################################################################

#reindex the matrix and separtate the data

#first calculate the index number and then add the index to the data

#reorganize the data after that using rbind function

#####################################################################

rownum <- dim(KUZ)[1]

empty <- seq(rownum,1)

KUZT <- KUZ[empty,]

rownames(KUZT) <- 1:nrow(KUZT)

#####################################################################

#Organize the data according to the property

#Name it and reorganize it

#####################################################################

KUZ\_trade <- cbind(KUZT[,1:2], KUZT[,8:9], KUZT[,15:16], KUZT[,22:23], KUZT[,29:30], KUZT[,36:37], KUZT[,43:44], KUZT[,50:51])

KUZ\_bid <- cbind(KUZT[,3:4], KUZT[,10:11], KUZT[,17:18], KUZT[,24:25], KUZT[,31:32], KUZT[,38:39], KUZT[,45:46], KUZT[,52:53])

KUZ\_ask <- cbind(KUZT[,5:6], KUZT[,12:13], KUZT[,19:20], KUZT[,26:27], KUZT[,33:34], KUZT[,40:41], KUZT[,47:48], KUZT[,54:55])

l = dim(KUZT)[1]

trade1 <- data.frame(index0(KUZ\_trade[1169:l,1:2]))

trade2 <- data.frame(index0(KUZ\_trade[1097:l,3:4]))

trade3 <- data.frame(index0(KUZ\_trade[757:l,5:6]))

trade4 <- data.frame(index0(KUZ\_trade[529:l,7:8]))

trade5 <- data.frame(index0(KUZ\_trade[433:l,9:10]))

trade6 <- data.frame(index0(KUZ\_trade[313:l,11:12]))

trade7 <- data.frame(index0(KUZ\_trade[1209:l,13:14]))

trade8 <- data.frame(index0(KUZ\_trade[609:l,15:16]))

bid1 <- data.frame(index0(KUZ\_bid[1169:l,1:2]))

bid2 <- data.frame(index0(KUZ\_bid[913:l,3:4]))

bid3 <- data.frame(index0(KUZ\_bid[641:l,5:6]))

bid4 <- data.frame(index0(KUZ\_bid[377:l,7:8]))

bid5 <- data.frame(index0(KUZ\_bid[261:l,9:10]))

bid6 <- data.frame(index0(KUZ\_bid[,11:12]))

bid7 <- data.frame(index0(KUZ\_bid[1069:l,13:14]))

bid8 <- data.frame(index0(KUZ\_bid[385:l,15:16]))

ask1 <- data.frame(index0(KUZ\_ask[1169:l,1:2]))

ask2 <- data.frame(index0(KUZ\_ask[949:l,3:4]))

ask3 <- data.frame(index0(KUZ\_ask[637:l,5:6]))

ask4 <- data.frame(index0(KUZ\_ask[221:l,7:8]))

ask5 <- data.frame(index0(KUZ\_ask[253:l,9:10]))

ask6 <- data.frame(index0(KUZ\_ask[37,11:12]))

ask7 <- data.frame(index0(KUZ\_ask[305:l,13:14]))

ask8 <- data.frame(index0(KUZ\_ask[409:l,15:16]))

names(trade8) = names(trade7) = names(trade6) = names(trade5) = names(trade4) = names(trade3) = names(trade2) = names(trade1)

collection\_trade <- rbind(trade1, trade2, trade3, trade4, trade5, trade6, trade7, trade8)

names(bid8) = names(bid7) = names(bid6) = names(bid5) = names(bid4) = names(bid3) = names(bid2) = names(bid1)

collection\_bid <- rbind(bid1, bid2, bid3, bid4, bid5, bid6, bid7, bid8)

names(ask8) = names(ask7) = names(ask6) = names(ask5) = names(ask4) = names(ask3) = names(ask2) = names(ask1)

collection\_ask <- rbind(ask1, ask2, ask3, ask4, ask5, ask6, ask7, ask8)[,1:2]

#####################################################################

#Convert the data in to POSIXct so that we can do our next steps

#again here we add the weird number 2000\*31556926 in order to adjust #the time

#####################################################################

collection\_trade$Date <- as.POSIXct(collection\_trade$Date, origin = "11-00-2017 00:00:00 GMT", format = "%m/%d/%Y %H:%M", tz = "GMT")

collection\_trade$Date <- collection\_trade$Date+31556926\*2000+52000

collection\_bid$Date <- as.POSIXct(collection\_bid$Date, origin = "11-00-2017 00:00:00 GMT", format = "%m/%d/%Y %H:%M", tz = "GMT")

collection\_bid$Date <- collection\_bid$Date+31556926\*2000+52000

collection\_ask$Date <- as.POSIXct(collection\_ask$Date, origin = "11-00-2017 00:00:00 GMT", format = "%m/%d/%Y %H:%M", tz = "GMT")

collection\_ask$Date <- collection\_ask$Date+31556926\*2000+52000

#####################################################################

#Processing the missing data

#We use mean to fill in the blank

#Name function filling

#####################################################################

filling <- function(x){

y = vector()

row1 = x[1,2]

time = 1

for (i in 2:length(x$Date))

{

if (x$Date[i-1]==x$Date[i])

{

row1 <- row1+x[i,2]

time <- time+1

}else

{

row1 <- row1/time

y <- rbind(y,cbind.data.frame("Date" = x$Date[i-1],row1))

row1 <- x[i,2]

time = 1

}

}

y <- index0(y)

}

KUZ\_minute\_trade <- filling(collection\_trade)

KUZ\_minute\_bid <- filling(collection\_bid)

KUZ\_minute\_ask <- filling(collection\_ask)

#####################################################################

#Create a grid for our new data

#####################################################################

day1 = seq.POSIXt(from = as.POSIXct("11/22/2017 20:00",format = "%m/%d/%Y %H:%M"),to = as.POSIXct("11/23/2017 01:30",format = "%m/%d/%Y %H:%M"),by = "min")

day2 = seq.POSIXt(from = as.POSIXct("11/23/2017 19:30",format = "%m/%d/%Y %H:%M"),to = as.POSIXct("11/24/2017 01:30",format = "%m/%d/%Y %H:%M"),by = "min")

day3 = seq.POSIXt(from = as.POSIXct("11/26/2017 19:00",format = "%m/%d/%Y %H:%M"),to = as.POSIXct("11/27/2017 01:30",format = "%m/%d/%Y %H:%M"),by = "min")

day4 = seq.POSIXt(from = as.POSIXct("11/27/2017 19:00",format = "%m/%d/%Y %H:%M"),to = as.POSIXct("11/28/2017 01:30",format = "%m/%d/%Y %H:%M"),by = "min")

day5 = seq.POSIXt(from = as.POSIXct("11/28/2017 19:00",format = "%m/%d/%Y %H:%M"),to = as.POSIXct("11/29/2017 01:30",format = "%m/%d/%Y %H:%M"),by = "min")

day6 = seq.POSIXt(from = as.POSIXct("11/29/2017 19:00",format = "%m/%d/%Y %H:%M"),to = as.POSIXct("11/30/2017 01:30",format = "%m/%d/%Y %H:%M"),by = "min")

day7 = seq.POSIXt(from = as.POSIXct("11/30/2017 19:00",format = "%m/%d/%Y %H:%M"),to = as.POSIXct("12/01/2017 01:30",format = "%m/%d/%Y %H:%M"),by = "min")

day8 = seq.POSIXt(from = as.POSIXct("12/03/2017 19:00",format = "%m/%d/%Y %H:%M"),to = as.POSIXct("12/04/2017 01:30",format = "%m/%d/%Y %H:%M"),by = "min")

day = c(day1,day2,day3,day4,day5,day6,day7,day8)

lday = length(day)

KUZ\_timegrid= data.frame(Date = day, lastprice = rep(0,lday), bid = rep(0,lday), ask = rep(0,lday))

#####################################################################

#Create a new data frame by coping

#Multiply by two times for ask bid

#####################################################################

for (i in 1:length(KUZ\_timegrid$Date)){

for (j in 1:length(KUZ\_minute\_trade$Date)){

if(KUZ\_timegrid$Date[i] != KUZ\_minute\_trade$Date[j]){

#USDKRW\_timegrid[i,2:4] = 0.5\*(USDKRW\_minute[i,2:4]+USDKRW\_timegrid[i-1,2:4])

}else{

KUZ\_timegrid[i,2] <- KUZ\_minute\_trade[j,2]

}

}

}

for (i in 1:length(KUZ\_timegrid$Date)){

for (j in 1:length(KUZ\_minute\_bid$Date)){

if(KUZ\_timegrid$Date[i] != KUZ\_minute\_bid$Date[j]){

#USDKRW\_timegrid[i,2:4] = 0.5\*(USDKRW\_minute[i,2:4]+USDKRW\_timegrid[i-1,2:4])

}else{

KUZ\_timegrid[i,3] <- KUZ\_minute\_bid[j,2]

}

}

}

for (i in 1:length(KUZ\_timegrid$Date)){

for (j in 1:length(KUZ\_minute\_ask$Date)){

if(KUZ\_timegrid$Date[i] != KUZ\_minute\_ask$Date[j]){

}else{

KUZ\_timegrid[i,4] <- KUZ\_minute\_ask[j,2]

}

}

}

#####################################################################

#filling the missing data using the mean of data

#####################################################################

KUZ\_timegrid[2647,2:4] = KUZ\_timegrid[2646,2:4]

for (i in 2:length(KUZ\_timegrid$Date)-1){

if(KUZ\_timegrid[i,2] == 0){

if(KUZ\_timegrid[i+1,2] != 0 && KUZ\_timegrid[i-1,2] != 0){

KUZ\_timegrid[i,2] = 0.5\*(KUZ\_timegrid[i-1,2]+KUZ\_timegrid[i+1,2])

}else{

ran <- runif(1, -3, 3 )

if(KUZ\_timegrid[i+1,2] == 0 && KUZ\_timegrid[i-1,2] != 0){ #first of the day

KUZ\_timegrid[i,2] = KUZ\_timegrid[i-1,2] + ran

}else{

KUZ\_timegrid[i,2] = KUZ\_timegrid[i+1,2]

}

}

}

if(KUZ\_timegrid[i,3] == 0){

if(KUZ\_timegrid[i+1,3] != 0 && KUZ\_timegrid[i-1,3] != 0){

KUZ\_timegrid[i,3] = 0.5\*(KUZ\_timegrid[i-1,3]+KUZ\_timegrid[i+1,3])

}else{

ran <- runif(1, -3, 3 )

if(KUZ\_timegrid[i+1,3] == 0 && KUZ\_timegrid[i-1,3] != 0){ #first of the day

KUZ\_timegrid[i,3] = KUZ\_timegrid[i-1,3] + ran

}else{

KUZ\_timegrid[i,3] = KUZ\_timegrid[i+1,3]

}

}

}

if(KUZ\_timegrid[i,4] == 0){

if(KUZ\_timegrid[i+1,4] != 0 && KUZ\_timegrid[i-1,4] != 0){

KUZ\_timegrid[i,4] = 0.5\*(KUZ\_timegrid[i-1,4]+KUZ\_timegrid[i+1,4])

}else{

ran <- runif(1, -3, 3 )

if(KUZ\_timegrid[i+1,4] == 0 && KUZ\_timegrid[i-1,4] != 0){ #first of the day

KUZ\_timegrid[i,4] = KUZ\_timegrid[i-1,4] + ran

}else{

KUZ\_timegrid[i,4] = KUZ\_timegrid[i+1,4]

}

}

}

}

#####################################################################

#Return calculating

#####################################################################

KUZ\_return = data.frame(Date = day[-1], Trade\_return = rep(1,lday-1), Quote\_return = rep(1,lday-1))

for (i in 1:lday-1){

KUZ\_return[i,2] = (KUZ\_timegrid[i+1,2]-KUZ\_timegrid[i,2])/KUZ\_timegrid[i,2]

midP = (KUZ\_timegrid[i,3]+KUZ\_timegrid[i,4])/2

midN = (KUZ\_timegrid[i+1,3]+KUZ\_timegrid[i+1,4])/2

KUZ\_return[i,3] = (midN-midP)/midP

}

#####################################################################