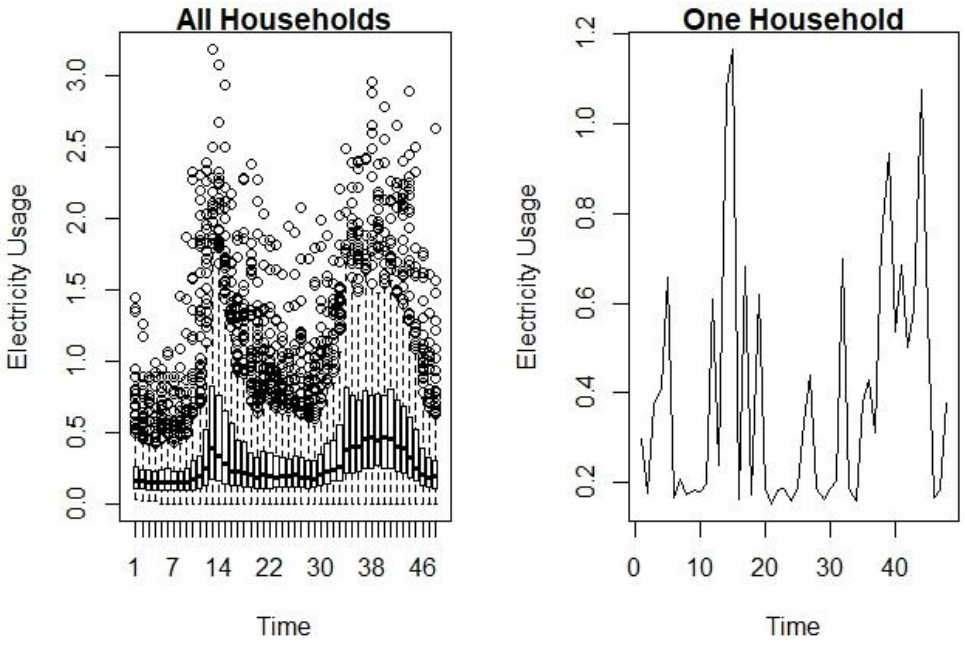
Jiawei Liu 9752650

QUESTION 1. ELECTRICITY MODELLING (60 MARKS)

1. Examine the boxplot and briefly discuss the aggregated pattern of electricity usage. Here is the boxplot produced by power.R:



We can find that in general there are two peaks or aggregated clusters in the boxplot. One is about 14, which means at 7am. The other is about 36 to 40, which means from 6pm to 8pm. This pattern is easy to understand. 7am is the time which most people get up and make breakfast while 6pm to 8pm is the time which most people make dinner and do some entertainment activities with family member after working or studying outside. Therefore, the usage of electricity is reasonably bigger in these two periods. Besides, amounts of usage in 12pm to 6am and 9am to 5pm are relatively smaller because those are times for sleeping and working/studying outside.

2. Construct a table of explanatory variables, where each row is a household, and each column is an explanatory variable that you have constructed from the original usage data. Describe the functions that you have constructed to create these explanatory variables, and in particular discuss how you have described (represented) the temporal structure of usage.

Here is the code to construct the table of explanatory variables:

```
simple.fn <- function(x,fn,cols=1:length(x))
{
```

```
    fn(x[cols])
     }

Build.Table <- function(p)
{
  tab <- apply(p,1,simple.fn,sum)
  tab <- cbind(tab,apply(p,1,simple.fn,var))
  tab <- cbind(tab,apply(p,1,simple.fn,sum,1:12))
  tab <- cbind(tab,apply(p,1,simple.fn,sum,13:24))
  tab <- cbind(tab,apply(p,1,simple.fn,sum,25:36))
  tab <- cbind(tab,apply(p,1,simple.fn,sum,37:48))
  tab <- cbind(tab,(apply(p,1,simple.fn,sum,1:12))/(apply(p,1,simple.fn,sum)))
  tab <- cbind(tab,(apply(p,1,simple.fn,sum,13:24))/(apply(p,1,simple.fn,sum)))
  tab <- cbind(tab,(apply(p,1,simple.fn,sum,25:36))/(apply(p,1,simple.fn,sum)))
  tab <- cbind(tab,(apply(p,1,simple.fn,sum,37:48))/(apply(p,1,simple.fn,sum)))

 colnames(tab) <- c("sum","var","sum1:12","sum13:24","sum25:36","sum37:48"
                    ,"percent1:12","percent13:24","percent25:36","percent37:48")
 as.data.frame(tab)
}
```

It can be seen that I construct 10 explanatory variables to distinguish the different overall patterns of usage. Actually, they can be grouped in 4 categories: sum, variance, 4 individual sums for 4 periods including 1:12 (12:30am-6am), 13:24 (6:30am-12am), 25:36 (12:30pm-6pm) and 37:48 (6:30pm-12pm), and 4 individual proportions for these 4 periods compared with the sum.

The sum can be used to distinguish two households with the same number of family members or rooms and different amounts of usage. It can also be helpful to classify different sizes of households because the big family with more members and rooms tends to use more electricity. But it may not always true.

The variance can be used to distinguish these two types of households. One is that its family members spend more time at home. Therefore, its pattern of usage may have some peaks at some periods and for the rest of time the amount of usage is also not too small. Its variance can be relatively smaller. The other is that its family members spend more time outside for studying and working. Therefore, its pattern of usage may have some peaks at some periods, such as breakfast time and dinner

time, and for the rest of time the amount of usage is reasonably small because only a few things keep consuming electricity without people at home, such as fridges and TPLINKs. Its variance can be relatively bigger.

Both 4 individual sums and 4 individual proportions for these 4 periods (1:12, 13:24, 25:36 and 37:48) can be used to distinguish these households which have different sizes and preference periods or habits to use electricity. The main reason why I construct both the sum and the proportions is that these 241 households can have considerably different size. Big households can use more electricity in periods 2 (13:24 (6:30am-12am)) compared with small households but it is just because they have more family members rather than preference or habits. Therefore, in general the sum is main for distinguishing different sizes while the proportion is main for distinguishing different habits or preferences.

3. Apply k-means clustering with 9 centers to the final explanatory data table. Produce two figures: one with boxplots showing the final clustering patterns of usage for each of the nine clusters, the second as 9 line plots showing the mean usage pattern for each timestep for each cluster.
Here is the code I use to plot these two figures:

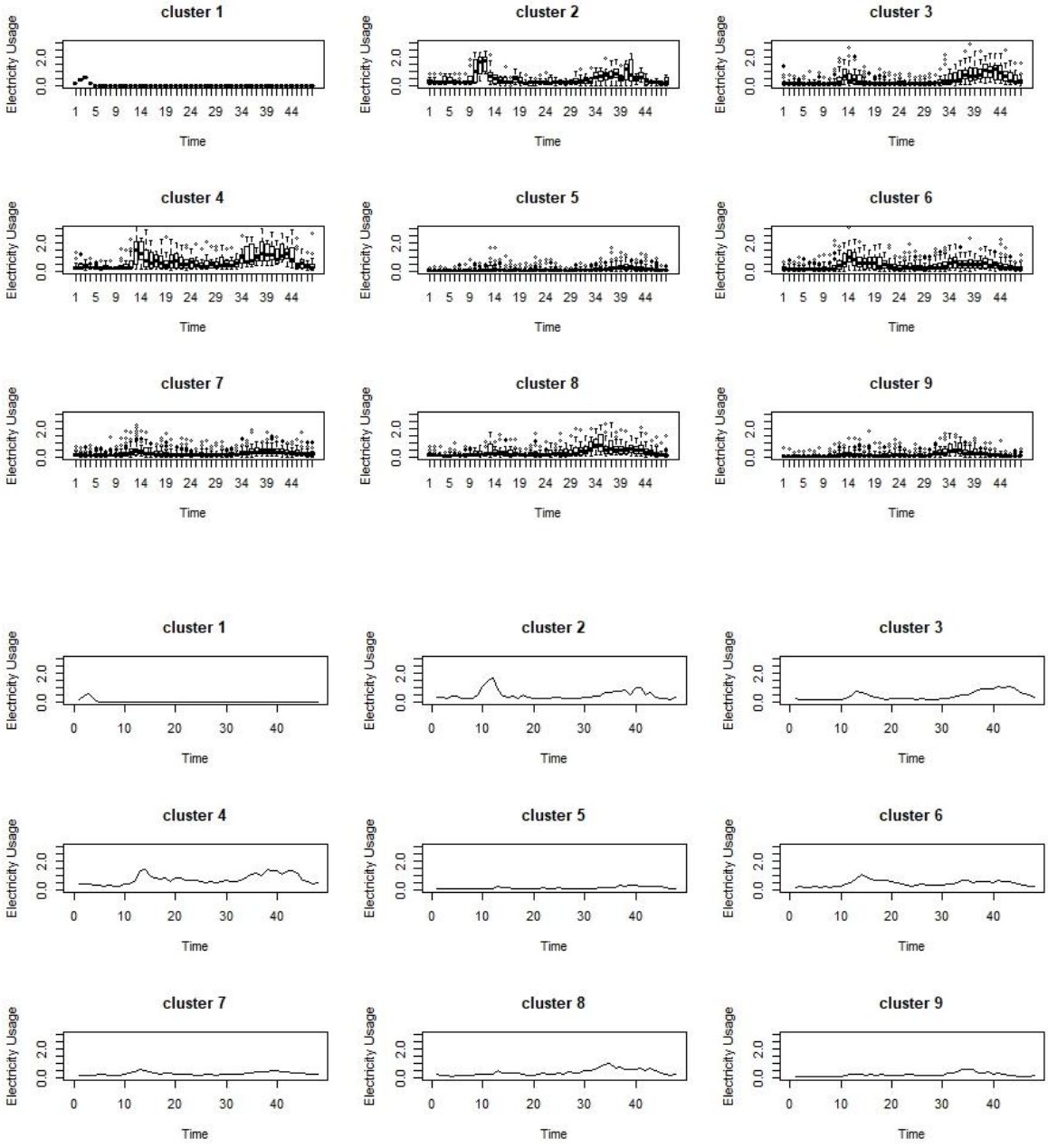
```
T <- Build.Table(p)
C <- do.cluster(T,9)

p <- as.data.frame(cbind(p,C$cluster))

par(mfrow=c(3,3))
for (i in 1:9) {
 f <- p[which(C$cluster==i),1:48]
 boxplot(f,xlab="Time",ylab="Electricity Usage", main=paste("cluster",i),ylim = c(0,3))
      }

  for (i in 1:9) {
     f <- p[which(C$cluster==i),1:48]
     f <- rbind(f,colMeans(f))
     plot(x=1:48,y=f[nrow(f),],xlab="Time",ylab="Electricity Usage",
     main=paste("cluster",i),xlim = c(0,48), ylim = c(0,3), type = "l")
      }
par(mfrow=c(1,1))
```

These two figures are shown below:



4. Comment on how the patterns of usage vary in the 9 clusters.

   It can be seen that in cluster 1 households only use low amount of power from 12:30am to 2am then they leave the houses and turn off everything. They may go out for a long holiday. In cluster 2 households use high amount of power in the morning and moderate or high amount of power in the evening while in the rest of time the amount is reasonably low. These two peaks may be due to breakfast, dinning and evening entertainment and people may be go out or sleep for the rest of time and only some machines keep working, like fridges. Cluster 3 is quite similar to cluster 2, and the difference is that for cluster 3 the amount of usage in

morning is relatively smaller, which is maybe because these households in clusters 3 do not make breakfast at home or not all family members have breakfast at home.

Cluster 4 is also similar to cluster 2. The difference is that for cluster 4 the amount of usage from 8am to 4pm is reasonably bigger, which might be because there are still some people stay at home and watch TV at that time. In cluster 5 households only use tiny amount of power in morning and evening. This type of households may be that some young people live alone and have breakfast and dinner outside rather than cooking themselves. In cluster 6 households use high amount of power in the morning and moderate amount of power in the evening. For these households, most of family members may have breakfast at home while only part of them have dinner at home and rest of them may eat outside.

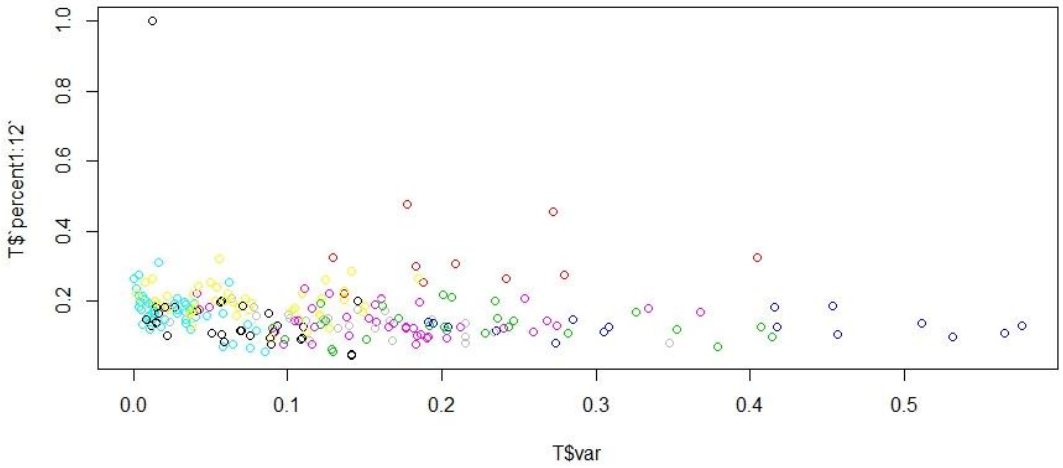
Cluster 7 is quite similar to cluster 5. The difference is that for cluster 7 the amounts of usage in morning and evening is slightly bigger compared with 5. The reason for this might be that these solitary people in cluster 7 cook at home. In cluster 8 households use low amount of power in the morning and moderate or high amount of power in the evening. This might be because these households do not cook or spend much time at home in the morning while they do cook or have some activities in the evening. Cluster 9 is quite similar to cluster 8. The difference is that for cluster 9 the amount of usage in the evening is slightly smaller because of the size of family or habits.

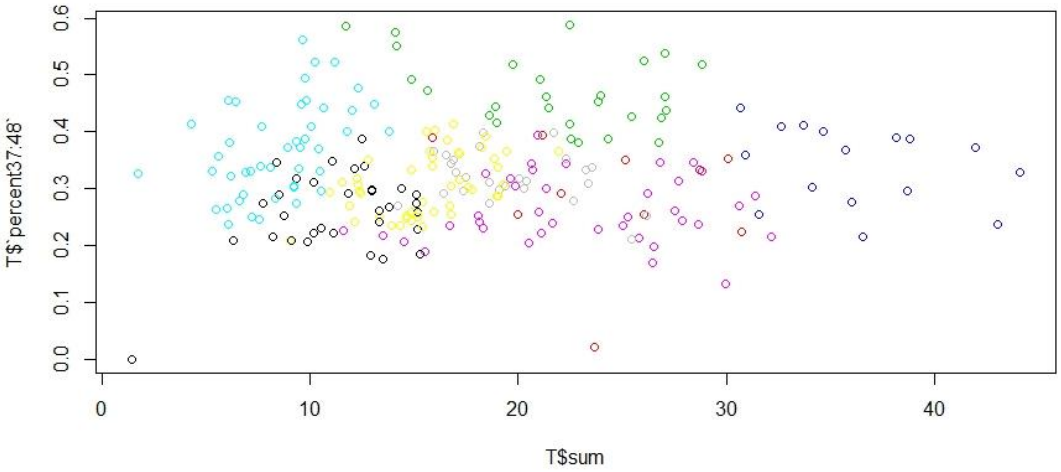
5. Select a pair of explanatory variables (say x1 and x2) that are not highly correlated and plot x1 versus x2 for each household colouring each point by cluster number. Comment on how the clustering is related to these variables.
   Here is the code I used to do this question:

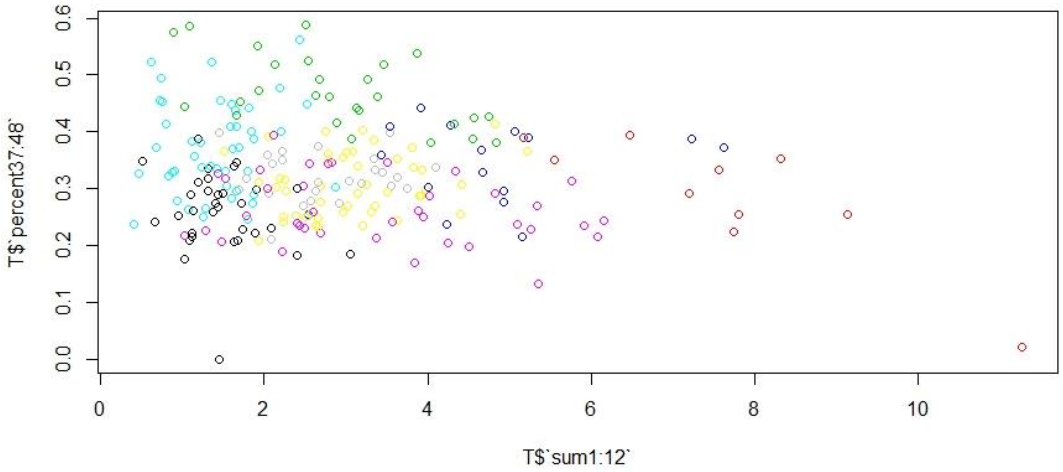
```
cor(T)
plot(T$var,T$`percent1:12`,col = C$cluster)
plot(T$sum,T$`percent37:48`,col = C$cluster)
plot(T$`sum1:12`,T$`percent37:48`,col = C$cluster)
```

   Firstly, I checked the correlation coefficients between all variables, and then three pairs of explanatory variables which are not highly correlated are selected: var and percent1:12, sum and percent37:48, sum1:12 and percent37:48.

   These three point plots are shown below:







From these graphs we can find that in general these variables can be helpful to classify these data and do clusters. Most of points with the same colour tend to get closer with each other and points with different colours tend to separated. For

example, if we have a look at the plot sum versus percent37:48, we can find that pink points have moderate values for both sum and percent37:48. Green points have moderate values for sum and high values for percent37:48 while blue points have low values for sum and moderate or high values for percent37:48. Therefore, different households have different values for different explanatory variables. these different values can be used to do clustering.

References:

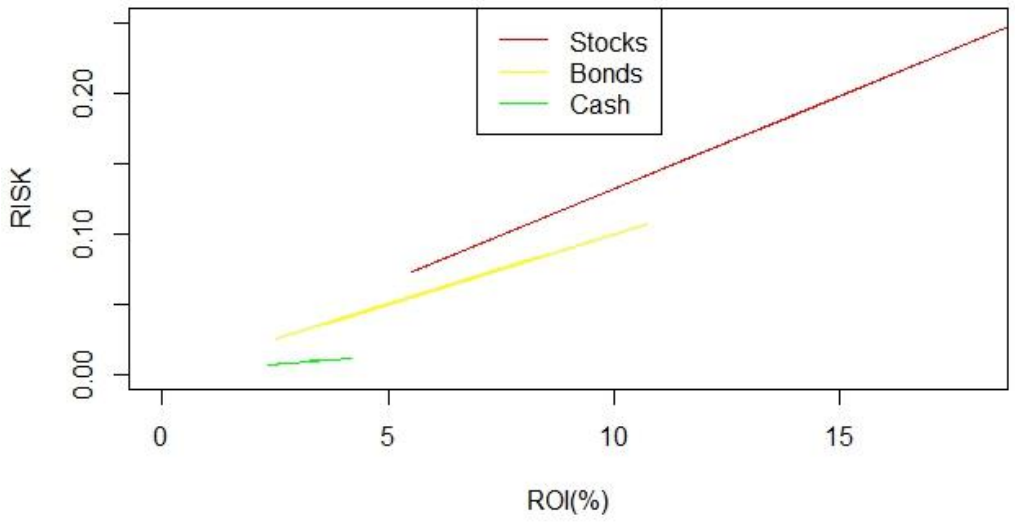
lecture 4 slides

QUESTION 2. INVESTMENT PORTOLIO MANAGEMENT (70 MARKS)

1. Visualise and discuss the different ROI and Risk associated with Stocks, Bonds and Cash.

   I produce a line plot to how ROI and Risk vary with regard to Stocks, Bonds and Cash. Here is the code and figure:

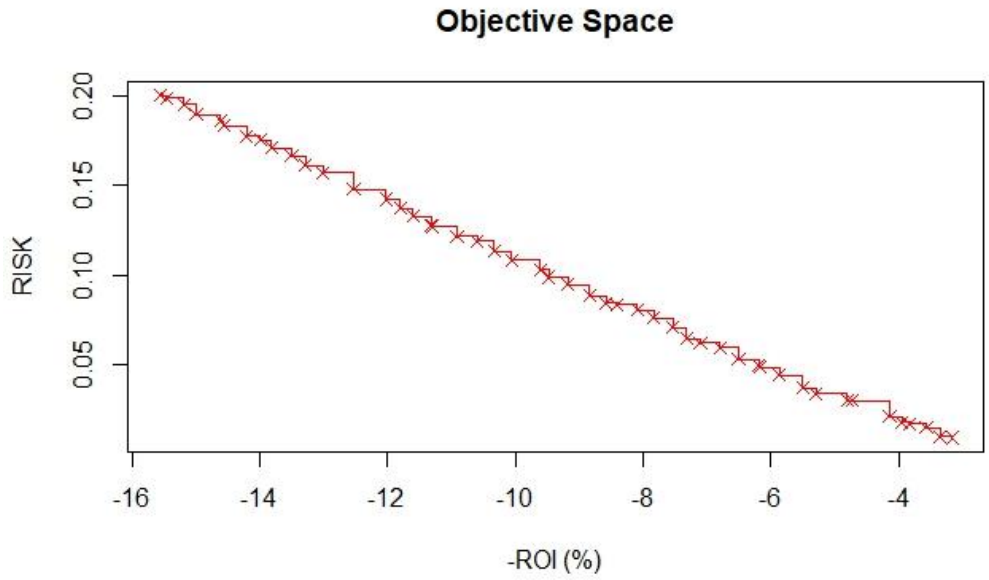
```
invest <- read.table("D:/Assn3/invest.tab")
a <- invest[1:10,]
b <- invest[11:20,]
c <- invest[21:30,]
plot(a$ROI,a$Risk,xlab = "ROI(%)",ylab = "RISK",xlim = c(0,18),ylim = c(0,0.25), type = "l",col = "red")
matlines(b$ROI,b$Risk,col = "yellow")
matlines(c$ROI,c$Risk,col = "green")
legend("top", legend = c("Stocks","Bonds","Cash"),
   col = c("red","yellow","green"), lty = 1)
```



We can find that in general these two variables are positive correlated. One increases and the other also increases. Cash has the smallest coefficient value which means that with the increasement of ROI Risk increases reasonably slowly. Besides both values and ranges of ROI and Risk for cash is small. Stocks have the biggest coefficient value here and it also has relatively big ranges and values of ROI and Risk. Finally, for bonds everything is moderate compared with cash and stocks.

2. Describe in words what the “invest.R” script is doing. In particular, state how a solution on the pareto front is represented and the relationship between the solution space, the objective space and the constraints.

   This script does a multi-objective criteria analysis to determine the best mix of stocks, bonds and cash over a range of trade-offs. Firstly, it sets that if anything smaller than 0.05, it will be treated as zero and max amount for a single option is 0.2. Secondly, it builds three constraints. The function portfolioSUM is to constrain that the sum of the portfolios must be between 0.95 and 1.0, which means that we should try to make use of all our money but the sum cannot be more than 1. The function portfolioRANGE is to constrain that each option must be bigger than 0.05 and smaller than 0.2, which means that we should not invest too much money to one option. The function portfolioNUM is to constrain that the total number of selected options should be bigger than 8 and smaller than 12, which means that we should not invest too many options while we should not also invest only a few options. Then, it sets two functions to calculate the sum of RISK and -ROI for each solution, which should be minimised (ROI should be maximised so here we use -ROI). Finally, portfolio is built by calling nsga2.



   The red line here is the pareto front of solution or solution space and all points on this front are non-dominates solutions. The solution is represented as a vector with 30 variables. If a variable is bigger than 0.05, it will be selected and none of them are bigger than 0.2. The upper right area of pareto front is search space which is full of dominates solutions while the left lower area of pareto front is infeasible space. The objective space is comprised of a list of ROI and a list of

RISK, which are calculated from solution space by function ROI and function RISK.

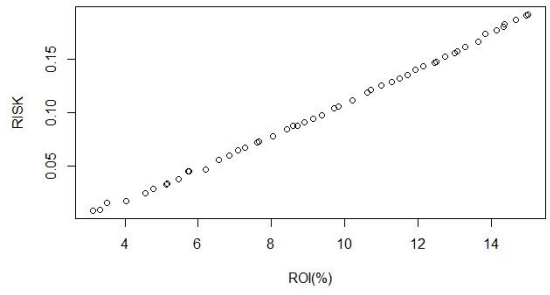
3. Using the result of the nsga2 model you have previously run, examine and present the blend of stocks, bonds and cash for a low risk, moderate risk and high-risk investment blend (just pick one from each general category). Discuss, in relation to Table 1, the level of risk that seems to be taken by the brokerage houses and whether the one-year return performance is related to the associated risk of the brokerage house.

   For this question, I produce a table called solution in which each row is a solution, and it has 5 columns. The first three cols are three proportions for cash, bonds and stocks respectively. The fourth col is the sum ROI for this solution and the fifth col is the sum Risk for this solution. Then I sort this table from the least to greatest risk and produce a table called vRISK. Here is the code and plot:

```
solution <- cbind(portfolio$par,-portfolio$value[,1],portfolio$value[,2])
for (x in 1:52){
  for (y in 1:30){
  if (solution[x,y] < 0.05)
  solution[x,y] = 0
  }
}

v <- rowSums(solution[,1:10])
v <- cbind(v,rowSums(solution[,11:20]))
v <- cbind(v,rowSums(solution[,21:30]))
v <- cbind(v,solution[,31])
v <- cbind(v,solution[,32])
colnames(v) <- c("Stocks","Bonds","Cash","ROI(%)","RISK")

vRISK <- as.data.frame(v[order(v[,5]),])
plot(vRISK$`ROI(%)`,vRISK$RISK,xlab = "ROI(%)",ylab = "RISK")
```



Here is part of the new table vRISK:

```
> show(vRISK)
          Stocks      Bonds       Cash    ROI(%)        RISK
 [1,] 0.00000000 0.00000000 0.95009699  3.130137 0.008943249
 [2,] 0.00000000 0.00000000 0.96904558  3.309106 0.009454587
 [3,] 0.05283917 0.00000000 0.89838436  3.511903 0.016373689
 [4,] 0.00000000 0.07710201 0.90008305  4.028737 0.017417004
 [5,] 0.00000000 0.15735171 0.81414075  4.567830 0.025104719
 [6,] 0.05124067 0.12694063 0.81907320  4.778395 0.028827079
 [7,] 0.05854617 0.12769471 0.80683409  5.132845 0.033444213
 [8,] 0.05854617 0.13370427 0.80228225  5.162102 0.033855353
 [9,] 0.05854617 0.15630347 0.77068610  5.471722 0.037776928
[10,] 0.10052125 0.24199768 0.61003467  5.719500 0.045344383
[11,] 0.10052125 0.24199768 0.61923517  5.758198 0.045454949
[12,] 0.06353796 0.26005377 0.66419082  6.195108 0.047393764
[13,] 0.15501737 0.15263506 0.68469166  6.554436 0.056231462
[14,] 0.15892500 0.16673068 0.64107039  6.847455 0.060072334
[15,] 0.19260066 0.16225602 0.60456439  7.084677 0.064740693
[16,] 0.20313293 0.16626673 0.60322983  7.278694 0.067205675
[17,] 0.25154201 0.12453136 0.61976722  7.623573 0.072406474
[18,] 0.25732285 0.12453136 0.61300244  7.664764 0.073217755
[19,] 0.22853173 0.19459320 0.55386410  8.044131 0.077924027
[20,] 0.22853173 0.19459320 0.55386410  8.044131 0.077924027
[21,] 0.28596436 0.18215569 0.52079486  8.419425 0.084665774
[22,] 0.24039195 0.30884275 0.41455951  8.589302 0.087950197
[23,] 0.26094165 0.15919811 0.56765587  8.710211 0.088192450
[24,] 0.27185759 0.16234901 0.55493052  8.904405 0.091043542
[25,] 0.29687111 0.19459320 0.49512148  9.136577 0.094607659
[26,] 0.27181532 0.26005377 0.45102935  9.378683 0.097527410
```

Then I pick three solutions from low risk, moderate risk and high risk respectively:

```
> vRISK[1,]
  Stocks Bonds      Cash   ROI(%)        RISK
1      0     0 0.950097 3.130137 0.008943249
> vRISK[26,]
       Stocks     Bonds      Cash   ROI(%)       RISK
26 0.2718153 0.2600538 0.4510294 9.378683 0.09752741
> vRISK[52,]
       Stocks     Bonds Cash   ROI(%)      RISK
52 0.7716258 0.2242119    0 15.01991 0.1917324
```

With regard to the Table 1, we can see that almost all brokerage houses choose the high risk investment blend as their solutions because they always want to prove that they can have higher ROI. When it comes to the relation between one year return performance and associated risk of the brokerage house, it can be concluded that the brokerage house with higher associated risk tends to have better one year return performance.

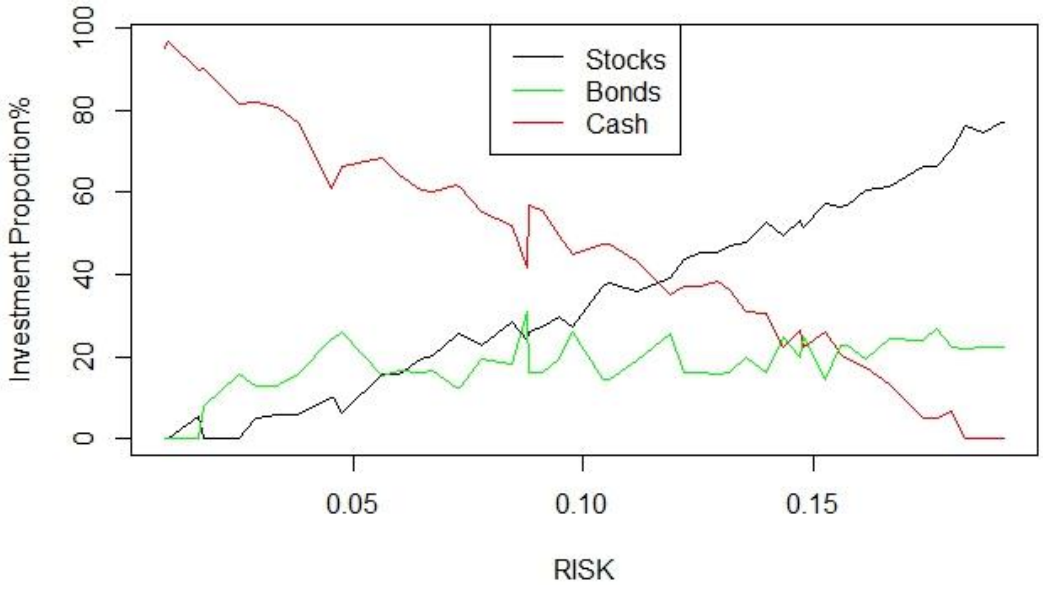
4. Examine the plot shown in Figure 3. This shows how the percentage of bonds, stocks and cash vary as you move along the pareto front from the least to greatest risk. Outline the approach (set of steps, algorithm, …) that would be required to produce this figure given the output from nsga2.

   I have already a table called vRISK in question 3. The vRISK has been in the order from the least to greatest risk. Then the plot can be easily produced. Here is the code for the first/left plot in Figure 3:

   vRISK <- as.data.frame(v[order(v[,5]),])

   plot(vRISK$RISK,100*vRISK$Cash,xlab = "RISK",ylab = "Investment

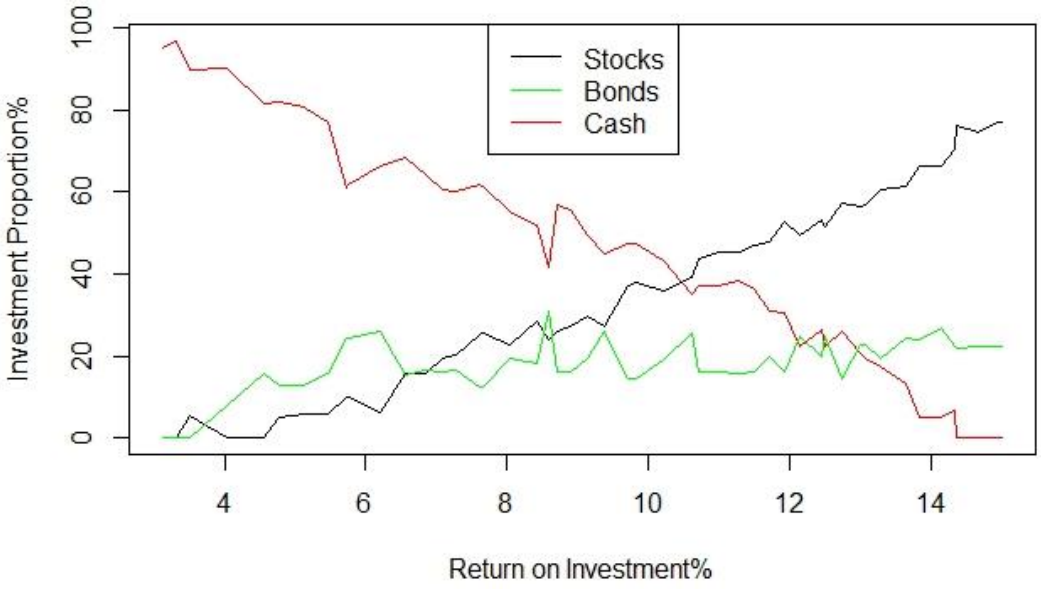
```
Proportion%"
        ,type = "l",col = "red")
    matlines(vRISK$RISK,100*vRISK$Stocks,col = "black")
    matlines(vRISK$RISK,100*vRISK$Bonds,col = "green")
    legend("top", legend = c("Stocks","Bonds","Cash"),
        col = c("black","green","red"), lty = 1)
```

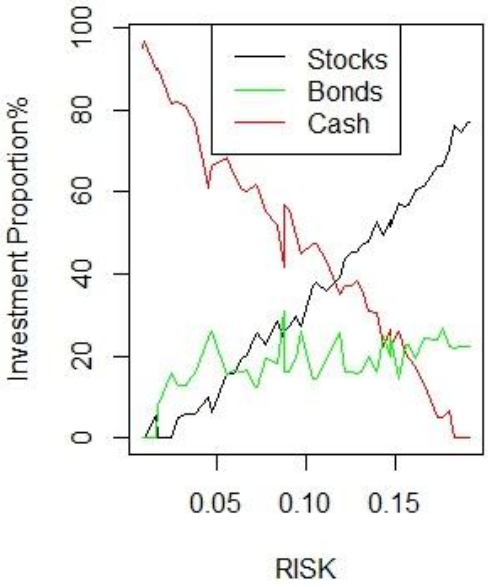


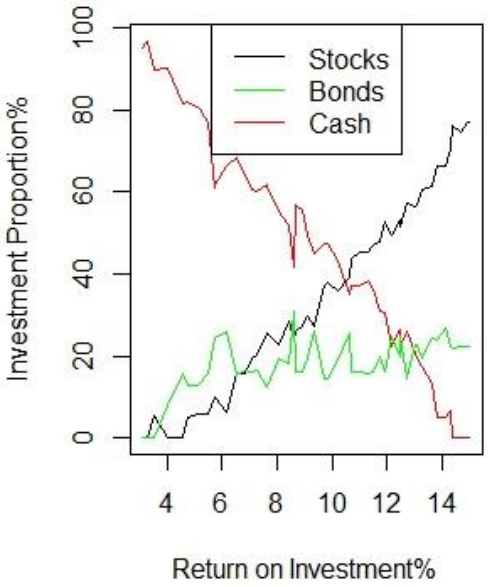
5. Produce one of the plots shown in Figure 3. Use the approach you have described in the previous question to comment and structure your R code. Include your "R" code and plot in your assignment.

   Here I just use the same approach in question 4 to produce the second/right plot in Figure 3. I build a table called vROI. The vROI has been in the order from the least to greatest ROI:

```
vROI <- as.data.frame(v[order(v[,4]),])
    plot(vROI$`ROI(%)`,100*vROI$Cash,xlab = "Return on
Investment%",ylab = "Investment Proportion%"
        ,type = "l",col = "red")
    matlines(vROI$`ROI(%)`,100*vROI$Stocks,col = "black")
    matlines(vROI$`ROI(%)`,100*vROI$Bonds,col = "green")
    legend("top", legend = c("Stocks","Bonds","Cash"),
        col = c("black","green","red"), lty = 1)
```







6. Find out some historical information on the performance of 2 of the riskiest brokerage houses from Table 1 and comment on whether they still exist, etc.
   It cannot obtain the exact numbers of risk for these brokerage houses. Lehman Brothers and Prudential may be the two riskiest brokerage houses because they recommended to invest most money to stocks which have the highest risk.

   Lehman Brothers was a global financial services firm and was the fourth-

largest investment bank in the United States, doing business in investment banking, equity and fixed-income sales and trading (especially U.S. Treasury securities), research, investment management, private equity, and private banking. Lehman was operational for 158 years from its founding in 1850 until 2008. The firm filed for Chapter 11 bankruptcy protection and do not exist anymore on September 15, 2008.

Prudential plc is a British multinational life insurance and financial services company headquartered in London, United Kingdom. It was founded in London in May 1848 to provide loans to professional and working people. Prudential has 26 million life customers. It owns Prudential Corporation Asia, which has leading insurance and asset management operations across 14 markets in Asia, Jackson National Life Insurance Company, which is one of the largest life insurance providers in the United States, and M&G Prudential, a leading savings and investments business serving customers in the UK and Europe. It still works well now and on July 20, 2017, Fortune Global 500 was released and Prudential Group ranked 56th.

7. Assume you only want to investigate the mix of solutions with the ROI between 9.0 and 13.0. State the constraint function that you would define to focus the search space between these ranges of ROI, and produce a plot of the pareto front using this additional constraint.
   For this question I add an additional constraint called portfolioROIRANGE, which is used to limit the range of ROI between 9.0 and 13.0. Here is the code and plot:

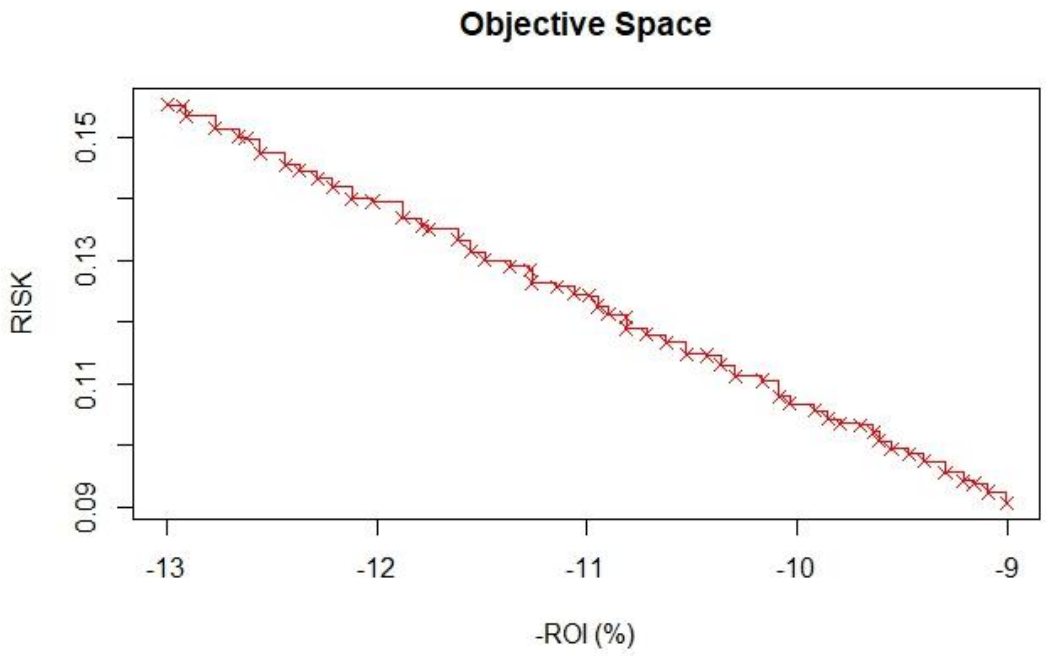
```
portfolioROIRANGE <- function(x)
{
  if (ROI(x) < -13.0) return(-1)
  if (ROI(x) > -9.0) return(-1)
  return(1)
}

constraint <- function(x)
{
  psum = portfolioSUM(x)
  prange = portfolioRANGE(x)
  pnum = portfolioNUM(x)
  pROIrange = portfolioROIRANGE(x)
```

```
    return(c(prange,pnum,psum,pROIrange))
}

portfolio2 <- nsga2(funs,idim=numberOptions,odim=2,
                    popsize=52,generations=1000,
                    lower.bounds=lower,upper.bounds=upper,
                    constraints = constraint,cdim=4)

plot(portfolio2,xlab="-ROI (%)",ylab="RISK",main="Objective Space")
```



References:
https://en.wikipedia.org/wiki/Lehman_Brothers
https://en.wikipedia.org/wiki/Prudential_plc
lecture 19,20 slides