#1

#我使用 random forest 配適模型

#先將 chr 型態的變數轉成 factor 代替原本的變數

#再將 Purchased=0, 1 的資料分別提取出來取 8:2 作為 training set 和 testing set

#利用 training set 做 random forest

#由 varImplot()的圖，因為 MeanDecreaseAccuracy 和 Gini 越大者對模型越重要，所以綜合來看我覺得比較明顯會影響 Purchased 的變數是 MSalary 和 Age

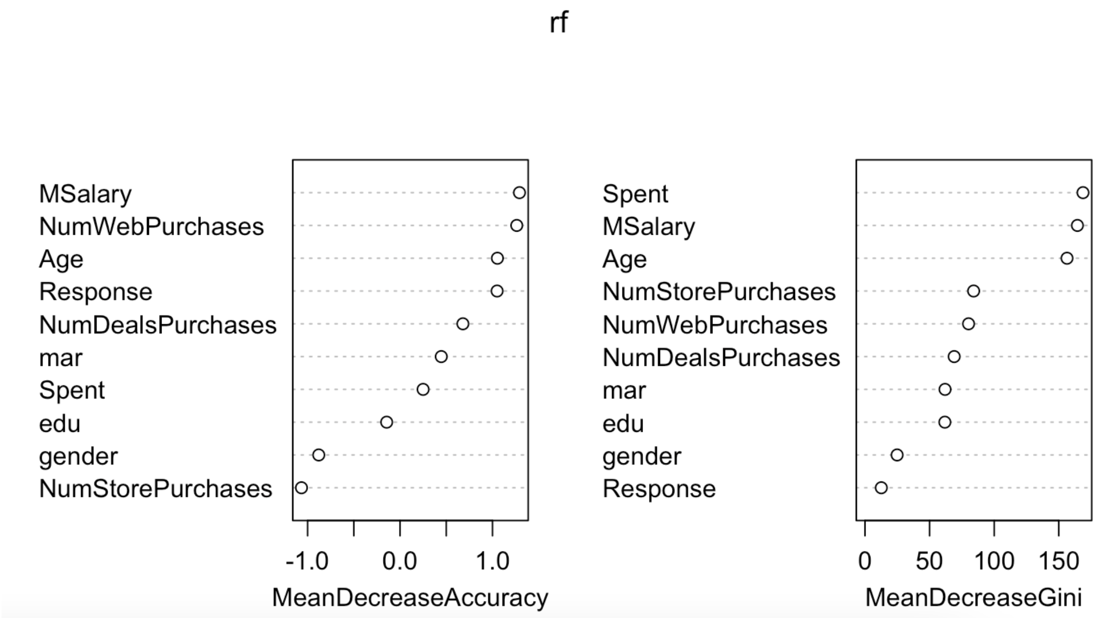
#對模型引入 testing set 做 predict，得到 Code 最後面 rfcm 之 table

Code:

```
library(tidyverse)
net <- read.csv("socialnetwork.csv")
net <-net[,-c(1)]#net2 第一欄為順序  不要
newnet <- net %>%
  mutate(
    edu=as.factor(ifelse(Education=='basic',0,ifelse(Education=='highschool',1,
      ifelse(Education=='college',2,ifelse(Education=='Master',3,4))))),
    mar=as.factor(ifelse(Marital_Status=='Absurd',0,ifelse(
      Marital_Status=='Alone',1,ifelse(Marital_Status=='Divorced',2,
      ifelse(Marital_Status=='Married',3,ifelse(Marital_Status=='Single',4,
      ifelse(Marital_Status=='Together',5,ifelse(Marital_Status=='Widow',6,7)))))))),
    gender=as.factor(ifelse(Gender=='Male',0,1)),
  )
newnet <-newnet[,-c(1,2,7)]
newnet$Response <- as.factor(newnet$Response)
newnet$Purchased <- as.factor(newnet$Purchased)

data0=newnet[newnet$Purchased==0,]
data1=newnet[newnet$Purchased==1,]

nrow0<-nrow(data0)
nrow1<-nrow(data1)

set.seed(3333)
train0=data0[sample(1:nrow0,0.8*nrow0),]
test0=data0[-sample(1:nrow0,0.8*nrow0),]

train1=data1[sample(1:nrow1,0.8*nrow1),]
test1=data1[-sample(1:nrow1,0.8*nrow1),]
```

#2

```
train=rbind(train0,train1)
rownames(train)<-1:nrow(train)

test=rbind(test0,test1)
rownames(test)<-1:nrow(test)
library(randomForest)
rf<-randomForest(Purchased ~.,train,ntree=100,importance=T)
plot(rf)
varImpPlot(rf)
```




```
pred=predict(rf,newdata = test)
rfcm<-table(Real=test$Purchased,Predict=pred)
rfcm
```

```
> rfcm
     Predict
Real   0   1
   0 204  18
   1  23 204
```

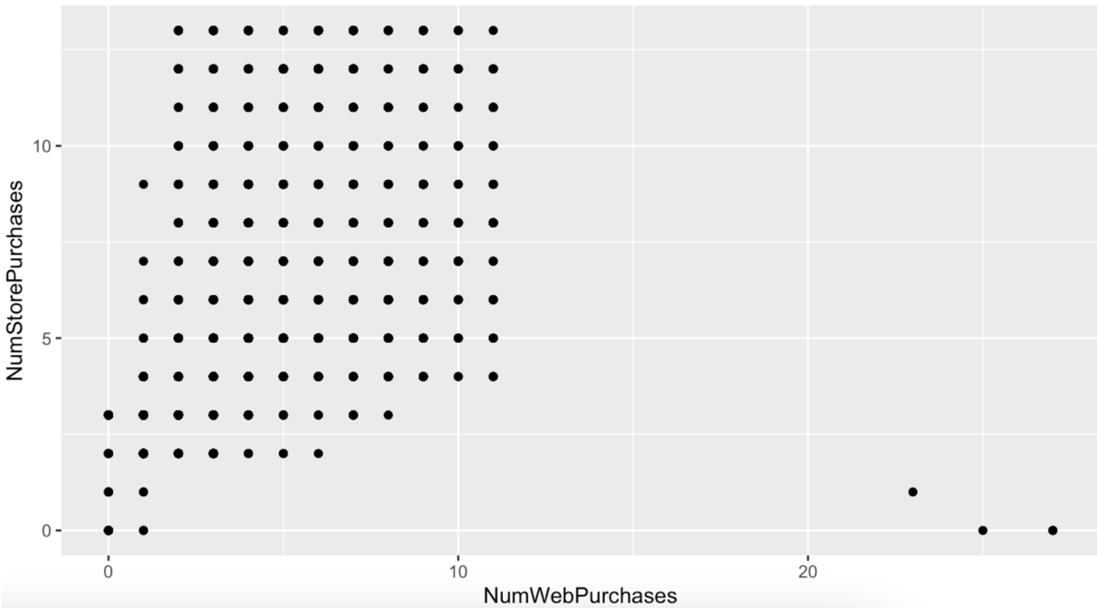
#2

#我使用 k-means 配適模型，並選擇我有興趣的變數 NumWebPurchases 和 NumStorePurchases，因為我覺得在網路上購買就比較不會在實體店面購買，反

之亦然。所以感覺可以有明顯的群可以分出來討論
#先將數字分佈範圍較大的 MSalary, Spent, Age 縮小至 0-10 之間
#可能是原始資料集本身的關係，幾乎每兩個變數的點分佈圖看起來都沒有明顯的分群，而觀察 NumWebPurchases 和 NumStorePurchases 的點分佈圖，雖然也看不太出明顯的分群但也不算分佈的非常平均，因此繼續使用這兩個變數做 k means
#利用 Elbow method 和 silhouette 方法尋找最佳的分群數大約都是 k=2
#分 2 群再用 fviz_cluster()得到最終分群結果
#雖然分群結果不甚理想，總之我歸納出藍色群顧客（基本上兩方面都很少購買）和紅色群顧客（網購店購都有但網購稍微多一點）。藍色群顧客因為少購買量所以我認為比較不用特別為他們制定策略；而紅色群顧客明顯較多，所以網購應該會是不少人的偏好，我會建議該公司設計更完善的網購 SOP、推行網購打折等策略吸引紅色群顧客

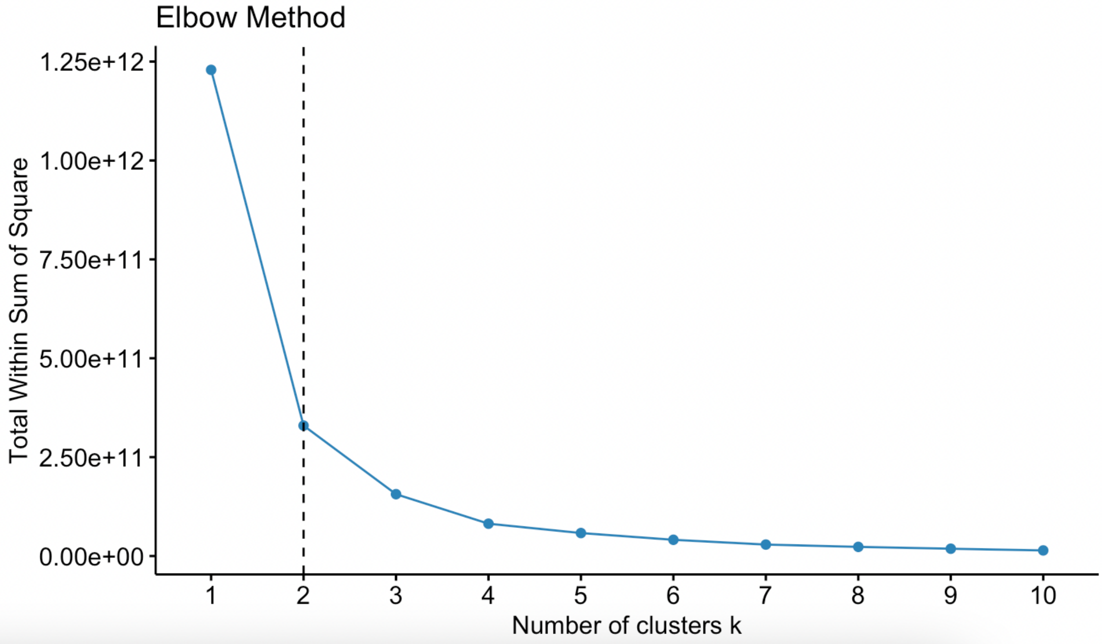
Code:

```
newnet2 <- newnet %>%
  mutate(
    MSalary = (MSalary - min(MSalary)) / (max(MSalary) - min(MSalary))*10,
    Spent = (Spent - min(Spent)) / (max(Spent) - min(Spent))*10,
    Age = (Age - min(Age)) / (max(Age) - min(Age))*10,
  )
ggplot(newnet2, aes(x=NumWebPurchases, y=NumStorePurchases)) +
  geom_point()
```

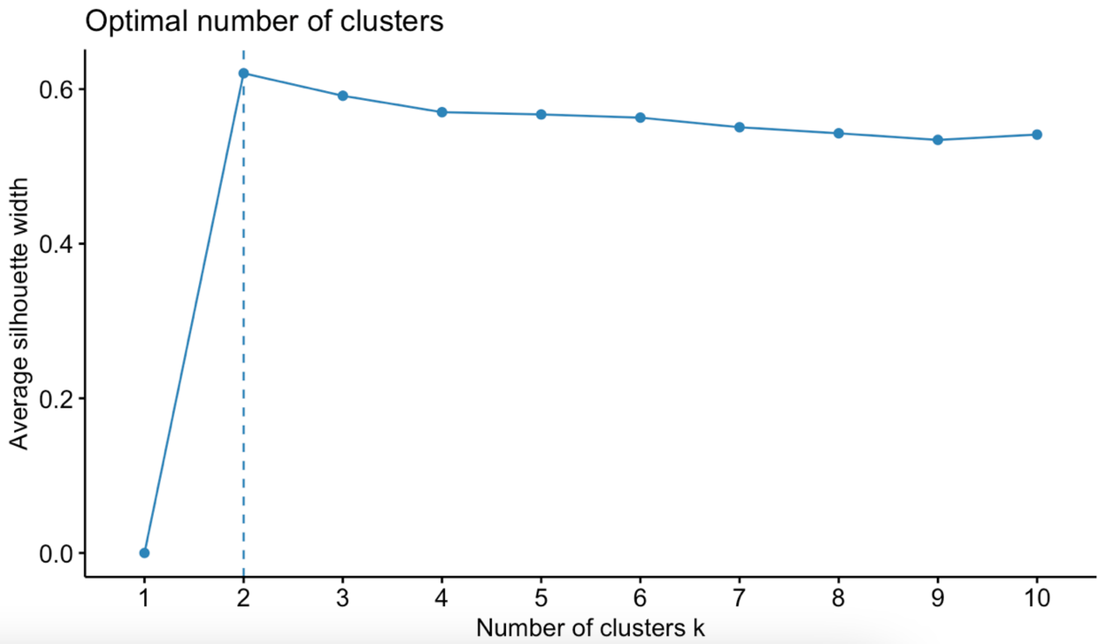



```
library(factoextra)
p = fviz_nbclust(newnet2,
                 FUNcluster = hcut,   # hierarchical clustering
```

```
                        method = "wss",          # total within sum of square
                        k.max = 10               # max number of clusters to consider
)
p
(p = p + labs(title="Elbow Method") )
p + geom_vline(xintercept = 2,            # elbow 在  X=2 的地方
                      linetype = 2)
```




```
fviz_nbclust(newnet[,c(1:11)], kmeans, method = "silhouette")
```




```
k = kmeans(newnet2[,c(3:4)], centers=2)#取 k=2
fviz_cluster(k,                   #  分群結果
```

```
data = newnet[,c(3:4)],                    #  資料
geom = c("point","text"), #  點和標籤(point & label)
ellipse.type = "norm")           #  框架型態
```