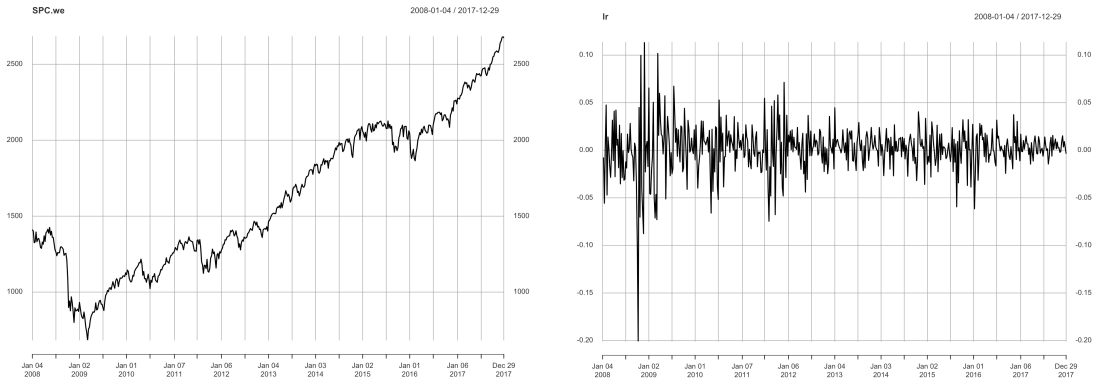
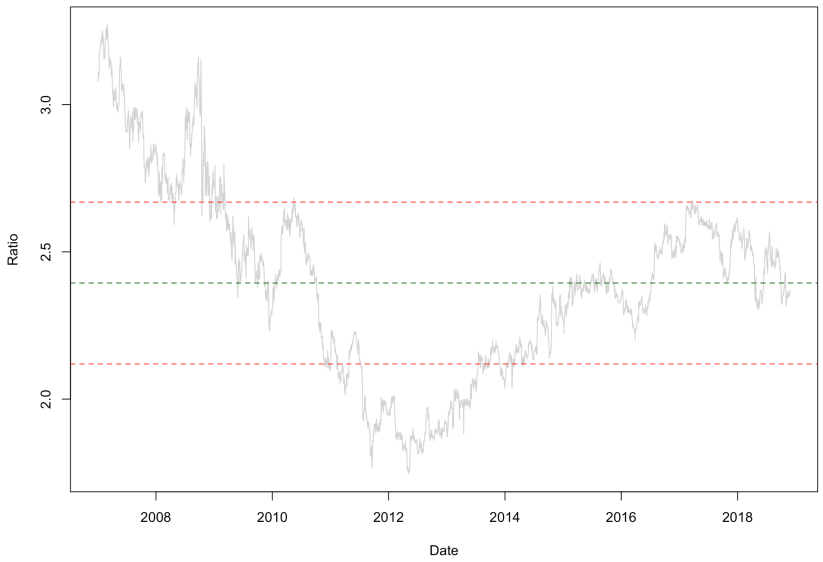
Problem 1

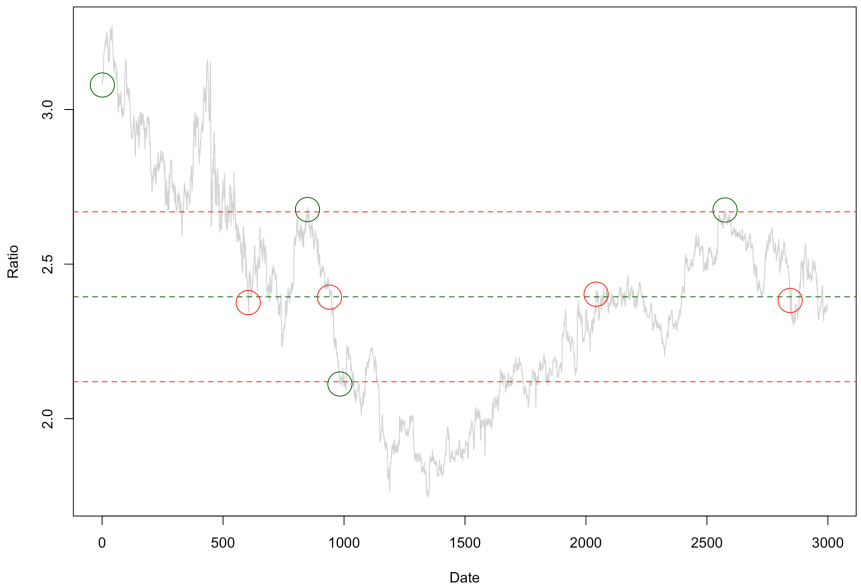
- *The symbols DPS download failed after two attempts. Error message: HTTP error 404.*



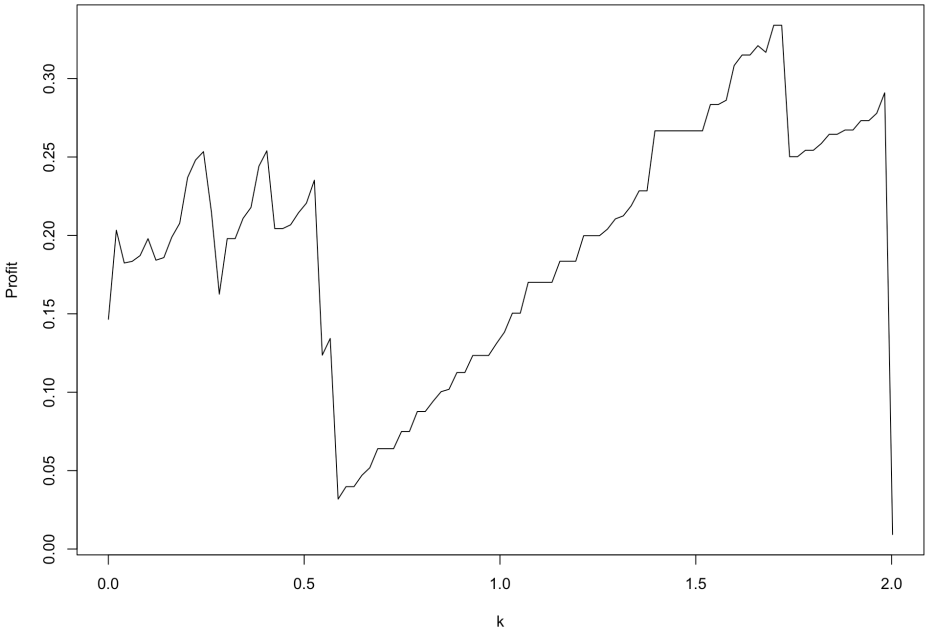
- *Plot ratio of two stocks（PepsiCo & Coca-Cola）*



- *Finding opening and Closing Positions*
  - *The green circles stand for the opening position*
  - *The red circles stand for the closing position*



- *Finding the optimal value for K based on the profit of different K and different positions.*
  - *split the ten years data as train and test*
  - *Compute the optimal K value using training data*
  - *After optimal k value to calculate the profit of testing data.*

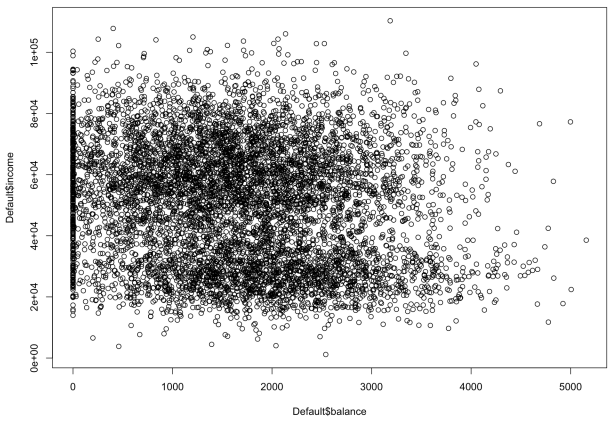


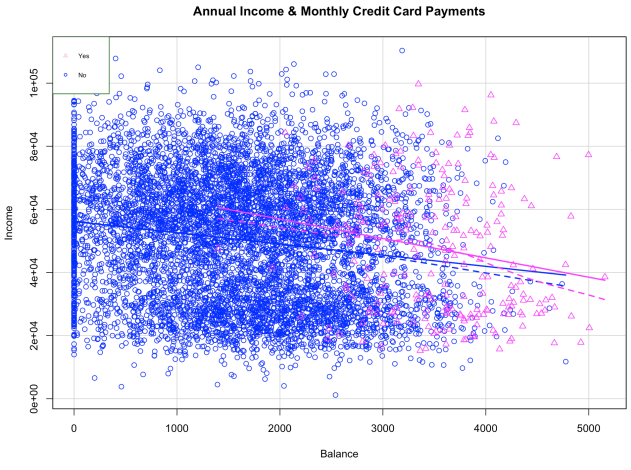

```
> ks[  profits == max(profits) ]
[1] 1.699071 1.719295

> testProfit
[1] 0.3946535
```

Problem 2

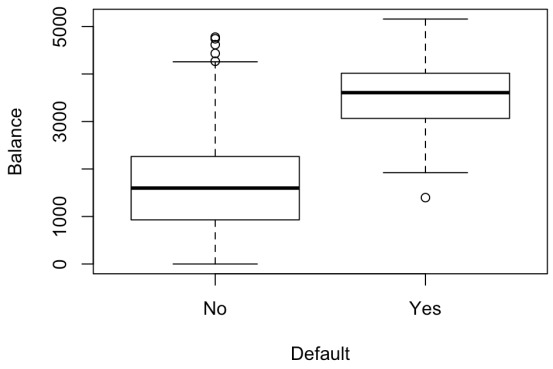
```
> summary(Default)
     index      default    student       balance           income
 Min.   :   1   No :6766   No :4956   Min.   :   0.0   Min.   :  1158
 1st Qu.:1751   Yes: 234   Yes:2044   1st Qu.: 956.8   1st Qu.: 32046
 Median :3500                         Median :1641.6   Median : 51961
 Mean   :3500                         Mean   :1671.2   Mean   : 50352
 3rd Qu.:5250                         3rd Qu.:2334.8   3rd Qu.: 65606
 Max.   :7000                         Max.   :5156.9   Max.   :110331
```

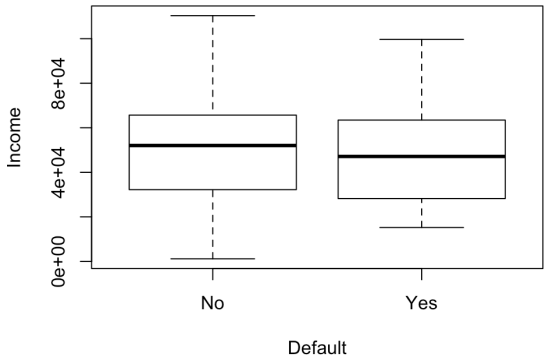
- *From the above summary we can found that this dataset is imbalanced dataset. Almost 97% of the dataset are No default, only 3% are default.*
- *We need to use confusion matrix to evaluate our model*



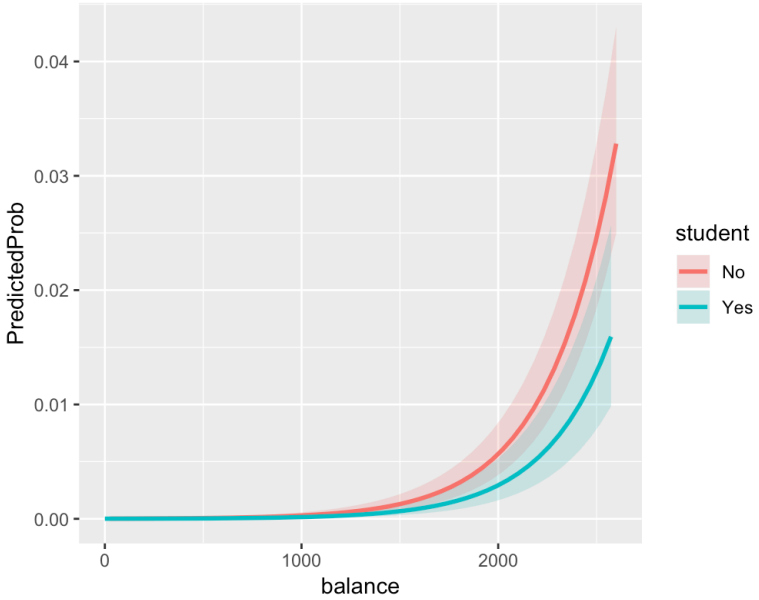


- *The first plot is quiet unclear to show the relationship between income and balance*
- *Check the relationship between balance and income how to effect the default which is a factor with levels No and Yes indicating whether the customer defaulted on their debt*
- *The second plot indicates that the higher average balance that the customer has remained on their credit card after making their monthly payment more likely the customers tends to have defaulted on their debt.*
- *For income, there is no obvious difference between the low income and high income customers.*





- *As above box-plots shows that higher the balance remained in their accounts the more likely the customers default the debt*
- *There is no obvious difference between the low income and high income customers whether they default the debt or not.*



Student is less risky than a non student with same credit balance

1. Logistic Regression

glm.pred	No	Yes
No	3373	74
Yes	19	35

-The overall error rate : 0.0266
-Error among defaulted : 0.679
-Sensitivity : is 0.321
-Specificity : 0.994

2. **Linear Discriminant Analysis**

lda.class	No	Yes
No	3381	82
Yes	11	27

- *Overall error rate : 0.0266*
- *Error among defaulted : 0.752*
- *Sensitivity : 0.248*
- *Specificity : 0.997*

3. **Quadratic discriminant Analysis**

qda.class	No	Yes
No	3378	80
Yes	14	29

- *Overall error rate : 0.0268*
- *Error among defaulted : 0.734*
- *Sensitivity : 0.266*
- *Specificity : 0.996*

4. **K-nearest Neighbor Classification**

knn1.pred	No	Yes
No	3285	84
Yes	107	25

- *Overall error rate : 0.055*
- *Error among defaulted : 0.771*
- *Sensitivity : 0.229*
- *Specificity : 0.968*

knn2.pred	No	Yes
No	3298	84
Yes	94	25

- *Overall error rate : 0.055*
- *Error among defaulted : 0.771*
- *Sensitivity : 0.229*
- *Specificity : 0.972*

knn3.pred	No	Yes
No	3367	97
Yes	25	12

- *Overall error rate : 0.034*
- *Error among defaulted : 0.890*
- *Sensitivity : 0.110*
- *Specificity : 0.993*

knn4.pred	No	Yes
No	3375	95
Yes	17	14

- *Overall error rate : 0.032*
- *Error among defaulted : 0.872*
- *Sensitivity : 0.128*
- *Specificity : 0.995*

knn5.pred	No	Yes
No	3387	107
Yes	5	2

- *Overall error rate : 0.032*
- *Error among defaulted : 0.982*
- *Sensitivity : 0.018*
- *Specificity : 0.999*

Conclusion:

For the dataset like problem 2 which are imbalance classification dataset. The overall error rate is not the proper model evaluation metrics.

- *Logistic Regression, LDA, and QDA has almost the same overall error rate and Specificity. However, the logistic regression is more accurate model. Since Logistic regression model have the highest Sensitivity which means the percentage of true defaulters been identified.*
- *AS for different K of K-NN algorithm. With K = 1 and 2, the KNN test error rate is the highest compare to other K values or other algorithms.*

Problem 3

a) Create a training set containing a random sample of 700 observations, and a test set containing the remaining observations.

```
set.seed(2)
train <- sample(1:nrow(OJ), 700)
oj_train <- OJ[train,]
dim(oj_train)
oj_test <- OJ[-train,]
dim(oj_test)
```

b) Fit a tree to the training data, with Purchase as the response and the other variables as predictors. Use the summary() function to produce summary statistics about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?

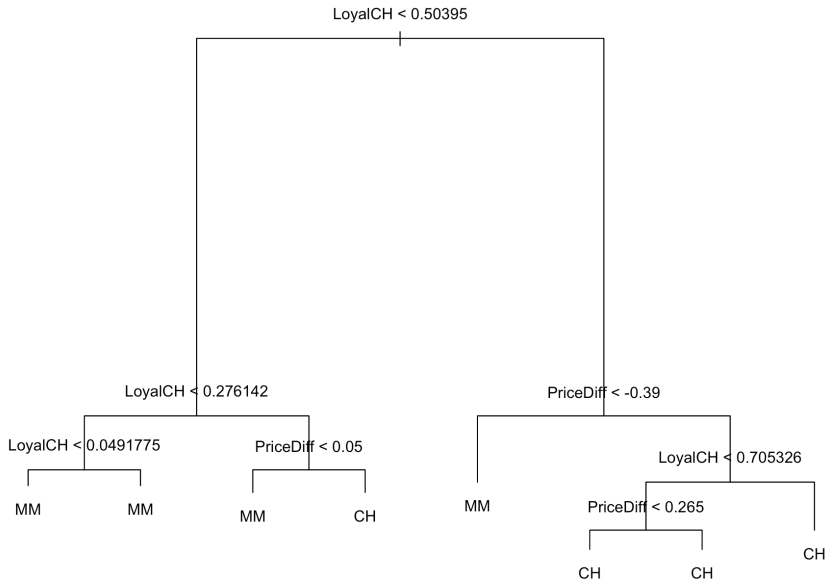
```
tree.OJ <- tree(Purchase~.,oj_train)
tree.OJ

tree.pred=predict(tree.OJ, oj_test ,type="class")
table(tree.pred, oj_test$Purchase)
summary(tree.OJ)

mean(tree.pred == oj_test$Purchase)
#overall error rate
1-mean(tree.pred == oj_test$Purchase)
```

```
> summary(tree.OJ)

Classification tree:
tree(formula = Purchase ~ ., data = oj_train)
Variables actually used in tree construction:
[1] "LoyalCH"   "PriceDiff"
Number of terminal nodes:  8
Residual mean deviance:  0.7493 = 518.5 / 692
Misclassification error rate: 0.1686 = 118 / 700
```

- *The model have 8 terminal nodes*
- *Training error rate is 0.1686*
- *Residual mean deviance is 0.7493*

c) Type in the name of the tree object in order to get a detailed text output. Pick one of the terminal nodes, and interpret the information displayed.

d) Create a plot of the tree, and interpret the results.



e) Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?

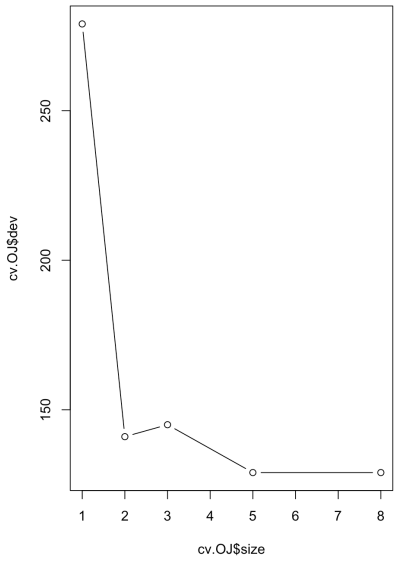
tree.pred	CH	MM
CH	210	42
MM	22	96

- *Test error rate: 0.172973*
- *Error among defaulted : 0.3043478*
- *Sensitivity : 0.6956522*
- *Specificity: 0.6862745*

f) Apply the cv.tree() function to the training set in order to determine the optimal tree size.

The optimal tree size is 5.

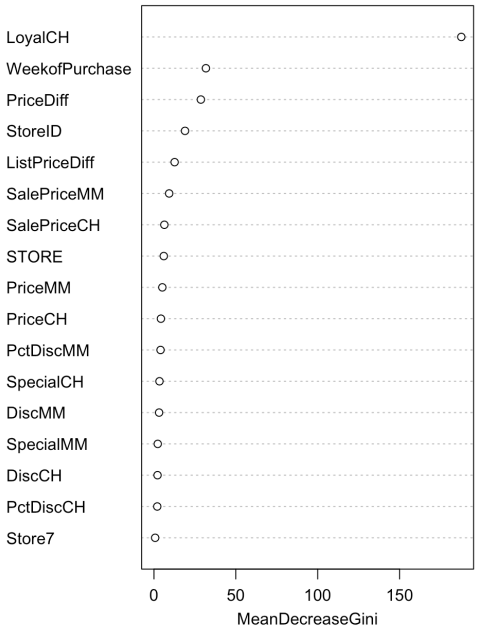
g) Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis.



h) Apply random forests method and check the performance of the model.

yhat.rf	CH	MM
CH	194	35
MM	38	103

- *Test error rate: 0.1972973*
- *Error among defaulted : 0.2463768*
- *Sensitivity : 0.7536232*
- *Specificity: 0.8362069*



i) Apply boosting method and check the performance of the model.

Have problem to implementing the gbm to classification problem.

j) Discuss the comparison in performance by applying different tree methods.

Random forest is more accuracy than the decision tree. Especially when we use confusion matrix to evaluate the model.