STAT462 ASSIGNMENT 3

(Chen Liang, Student ID 46275313)

Question2

(a)

Use set.seed(3), sample(nrow(carseat), 0.7*nrow(carseat)) to get the randomly split training set, and let the rest of the dataset to be the testing set. The obtained training set contains 280 observations and the testing set contains 120 observations.

(b)

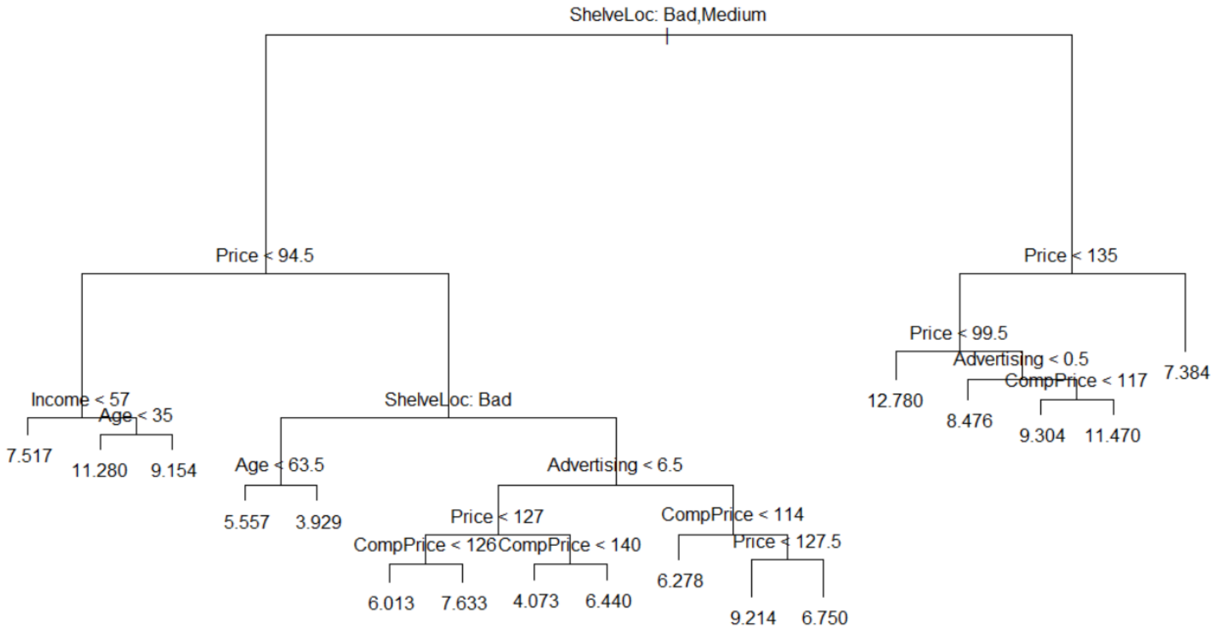


Figure 1 Regression Tree to the Training Set

Figure 1 shows the regression tree to the training set, from the tree, it could be checked that the variable 'ShelveLoc' is the most important when determining the responding variable 'Sales'. Since on the first node, the dataset is divided based on if the Location of Shelve is good, or bad or medium. At the same time, the branch with good location has a higher Sales than the branch with bad or medium one.

Table 1 displays the summary of the regression tree, which tells there are 6 variables being actually used when building the regression tree. The MSE for training set is 2.247 and that for testing set is 4.148.

Table 1 Summary of the Regression Tree

Variables actually used	"ShelveLoc" "Price" "Income" "Age" "Advertising" "CompPrice"
MSE for Trainset	2.247
MSE for Testset	4.148

(c)

It could be seen in Figure 2 that the size should be set as 5 because the deviation is the lowest. In this situation the MSE for testing set is 5.049. Therefore, pruning fails to improve the performance in this case.

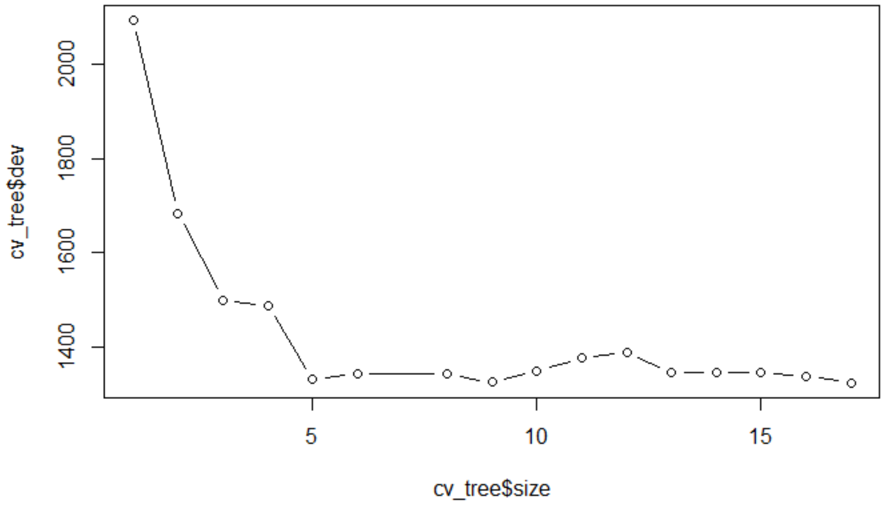


Figure 2 Relationship between Deviation and the Size of the Tree

(d)

When fitting the bagged regression tree, use m = p because bagged regression tree is the special situation of random forest at m = p. The MSE for training set is 0.483 and that for testing set is 0.647. Then fitting the random forest using sqrt(p) which is approximately equal to 3. The MSE for training set is 2.066 and that for testing set is 2.252. Since both the MSE for training and testing set do not decrease, decorrelating tree is not an effective strategy for this problem.

Table 2 Comparison of Bagged Regression and Random Forest

	mtry	MSE for training set	MSE for testing set
bagged regression	10	0.483	2.066
random forest	3	0.647	2.252

(e)

When fitting the model, set n.trees as 1000 to 5000 by step as 1000, interaction.depth ranging from 1 to 5, set shrinkage as 0.1, 0.01 and 0.001. The best model is the one with n.trees is equal to 3000, interaction.depth as 1 and shrinkage as 0.01.

(f)

The summary of Model 7 below shows that the most important variable is Price and ShelveLoc. The variable with the least importance is Urban and US.

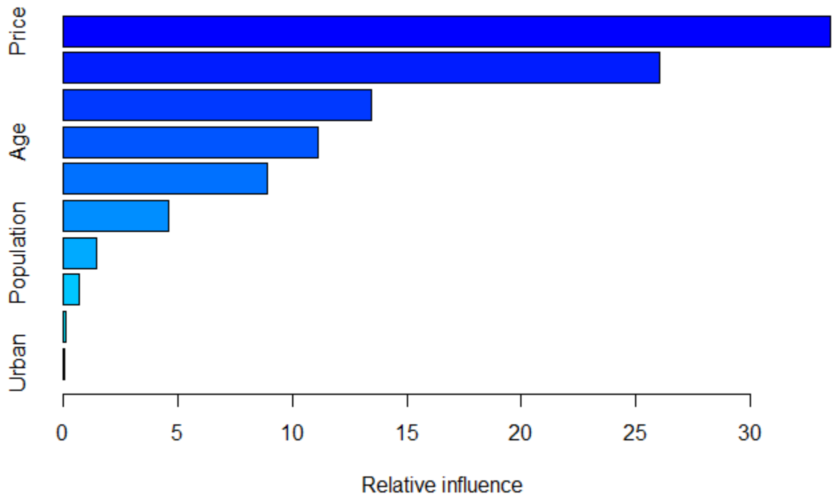


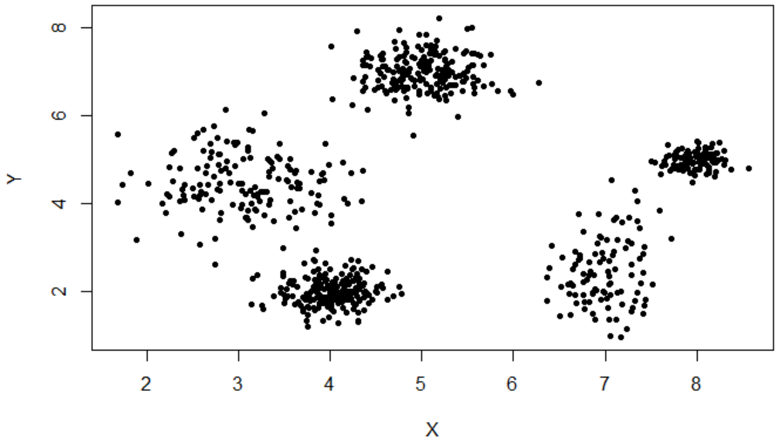
Table 2 Summary of the Model

variable	reLinf
Price	33.492
ShelveLoc	26.026
CompPrice	13.451
Age	11.146
Advertising	8.9325
Income	4.6079
Population	1.468
Education	0.714
US	0.0934
Urban	0.0695

Question3

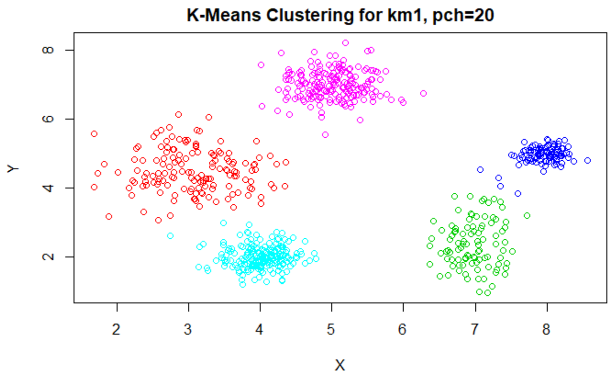
(a)

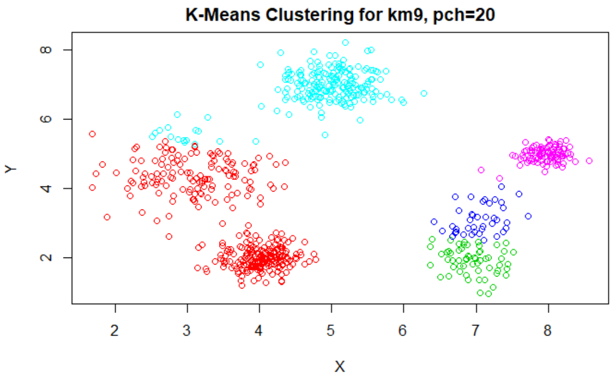
The data is displayed as below, and the number of clusters is estimated as 5 based on the figure.



(b)

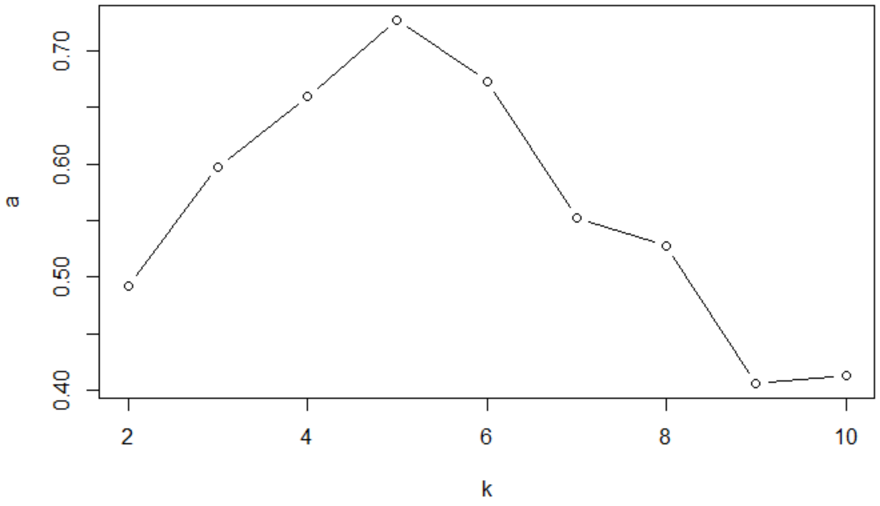
After ten times clustering, six different consequences are obtained. The best one and worst one are plotted as below.



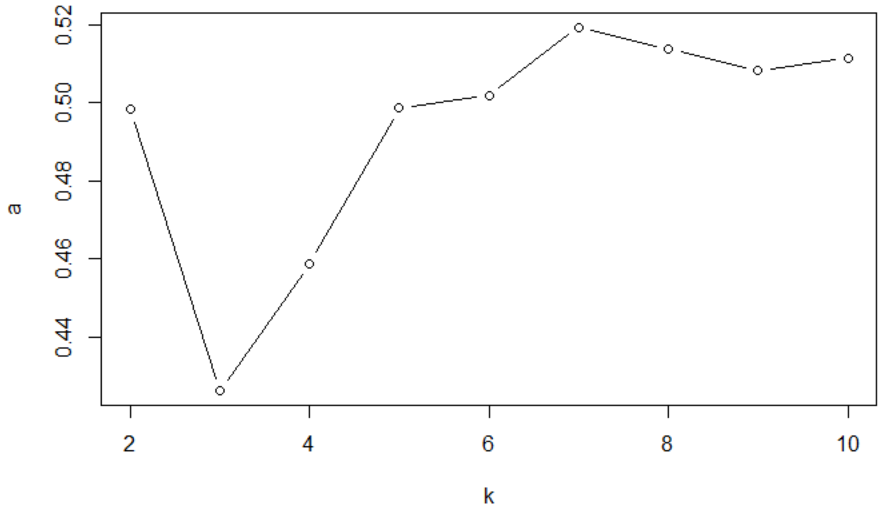


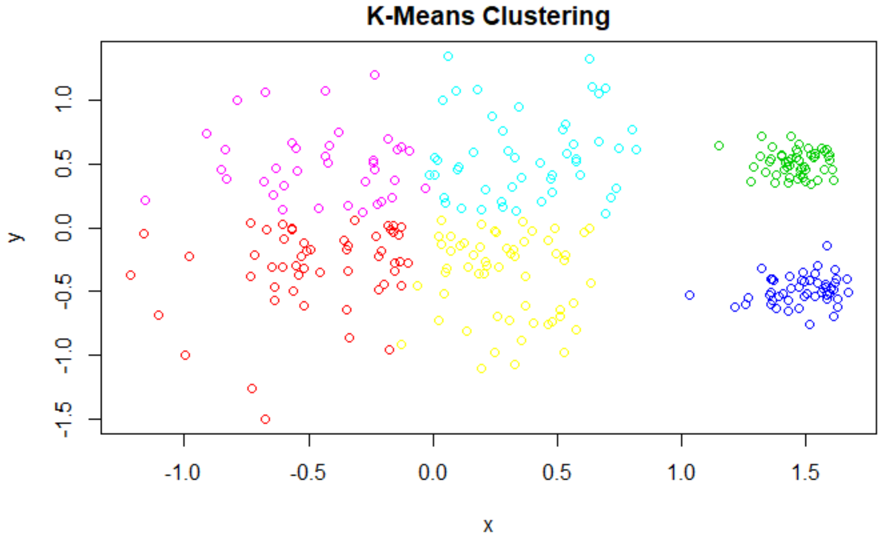
(c) Figure displays that the best k value for this data is 5 since average Silhouette coefficient reached the peak when k equals 5.

The silhouette coefficient is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette coefficient ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. Thus, higher average silhouette coefficient means better clustering. Therefore, k = 5 is recommended.



(d)



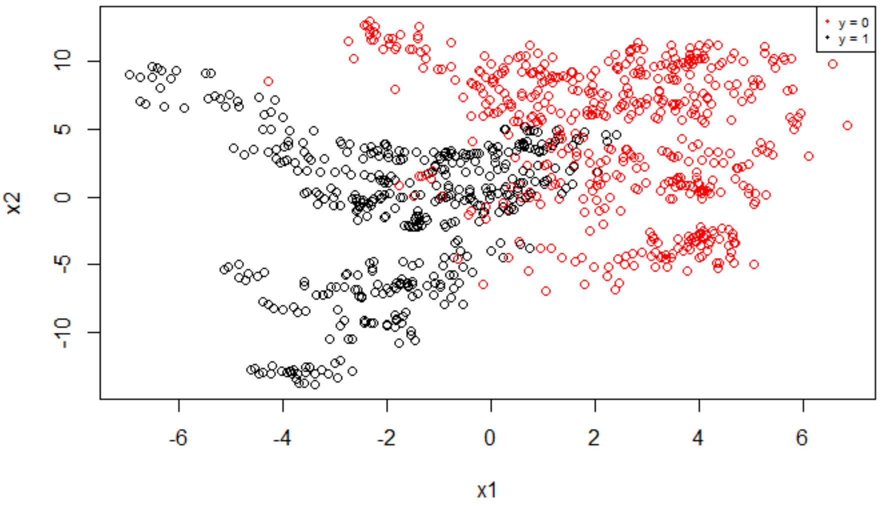


Question4

(a)

Use set.seed(1), sample(nrow(banknote), 0.7*nrow(banknote)) to get the randomly split training set, and let the rest of the dataset to be the testing set. The obtained training set contains 960 observations and the testing set contains 412 observations.

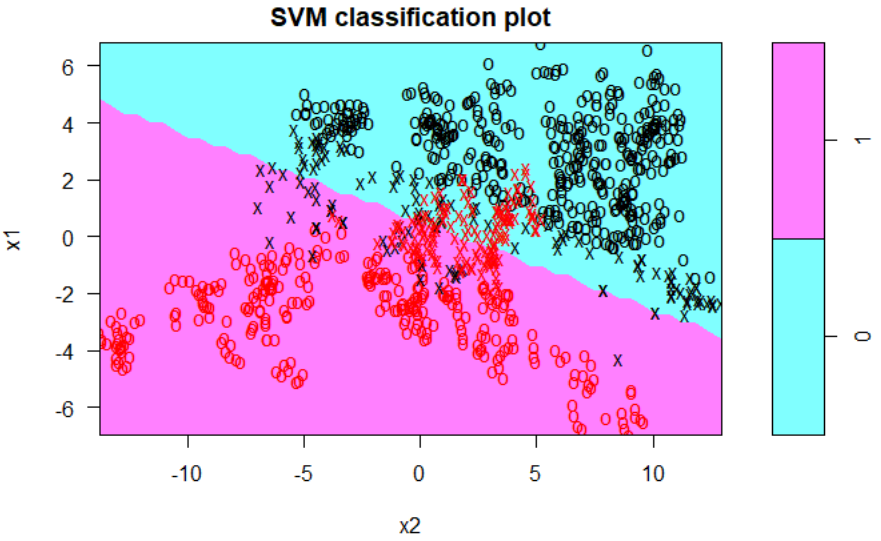
(b)



It could be seen clearly that it is not possible to find a separating hyperplane for the training data.

(c)

Fitting the model with cost = c(0.001, 0.01, 0.1, 1, 5, 10, 100).The best one resulting in the lowest error is the model with cost being equal to 0.1 and the model is plotted as below. From the summary of the best model, it could be figured out that since the cost I choose is low (0.1), which shows a greater tolerance of error and a higher bias will be produced. At the same time, the margine is wide and the number of support vectors is large which is 279.



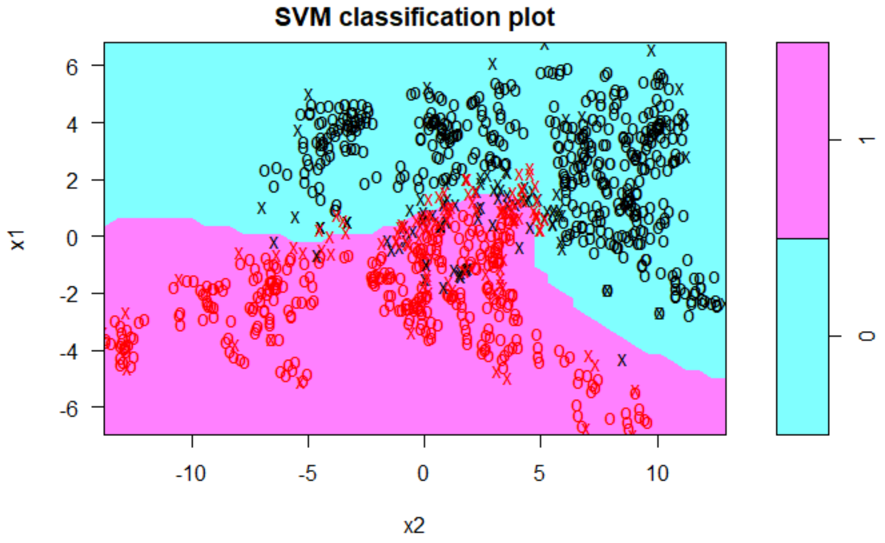
The contingency table tells that the accuracy rate is (199 + 168)/412 which is 89.29%. The serious error that the forged note being predicted to be genuine is 19/412which is 4.61%.

	truth

predict	0	1
0	199	19
1	26	168

(d)

The Model with cost 1 and gamma 2 performs best produces error as 0.0719 and get 207 support vectors. The value of cost is not that low as we get in (c), therefore, it would have a better balance in error and bias. At the same time, the value of gamma is not great so the overfitting could be effectively avoided.



The contingency table tells that the accuracy rate is (205 + 183)/(205 + 183 + 20 + 4) which is 94.17%. The serious error that the forged note being predicted to be genuine is 4/(05 + 183 + 20 + 4) which is 1.02%.

	truth	
predict	0	1
0	205	4
1	20	183