4.

(a)

如果 x<0.05,只能抽到[0,x+0.05],機率有(100x+5)%,積分得 0.375

同理,如果 x>0.95 只能抽到[x-0.05,1],機率有(105-100x)%,積分得 0.375

如果 x 介於[0.05,0.95]間,抽到[x-0.05,x+0.05]的機率都為 1 ,積分得 9

9+0.375+0.375=9.75

9.75%

(b)

機率為:0.0975^2=0.0950625

(c)

0.0975^100=7.951729e-102,趨近於 0

(d)

knn 是由 x 附近的點去估計,由 a,b,c 可見當維度越高的時候,越沒有附近的樣本點去作估計。

5. We now examine the differences between LDA and QDA.

(a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

(b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

(c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

(d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

(a)

QDA,LDA ，QDA 在 training set 上表現一定比 LDA 好,但是在如果真實資料近似線性的話,QDA 在 testing set 上表現會比 LDA 差 。這種狀況稱為 overfitting。

(b)

QDA,QDA, QDA 在 training set 上表現一定比 LDA 好,且如果如果真實資料是非線性的話,QDA 在 testing set 上表現也會比 LDA 好。

(c)

Improve,因為越多的資料可以避免 QDA 的 overfitting 的問題

(d)

False, QDA 在 training set 上表現可能比 LDA 好,但是在如果資料是可線性分割的話,QDA 在 testing set 上表現會比 LDA 差 。這種狀況稱為 overfitting。

6. Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.05, \hat{\beta}_2 = 1$.

   (a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

   (b) How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?

(a)

```
> p <- function(x1,x2)
+   { z <- exp(-6 + 0.05*x1 + 1*x2); return( round(z/(1+z),2))}
> p(40,3.5)
[1] 0.38
```

(b)

50hours

10. This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

(b) Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

(e) Repeat (d) using LDA.

(f) Repeat (d) using QDA.

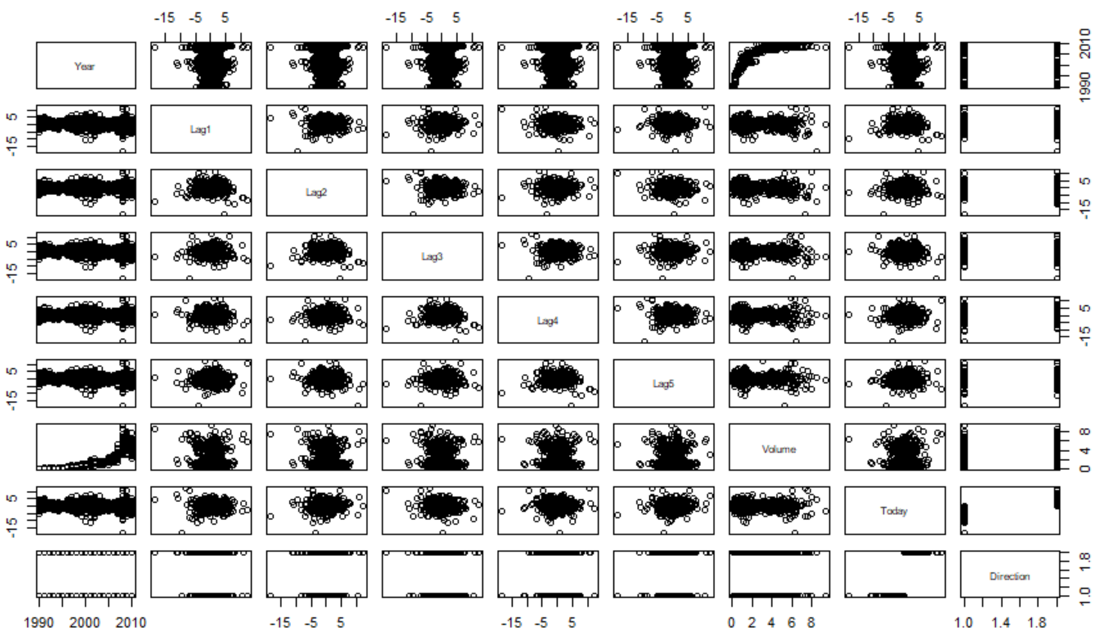
(g) Repeat (d) using KNN with $K = 1$.

(h) Which of these methods appears to provide the best results on this data?

(i) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for

(a)

```
> summary(Weekly)
      Year           Lag1               Lag2               Lag3               Lag4
 Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950
 1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580   1st Qu.: -1.1580
 Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410   Median :  0.2380
 Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472   Mean   :  0.1458
 3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090   3rd Qu.:  1.4090
 Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260
      Lag5              Volume            Today          Direction
 Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950   Down:484
 1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540   Up  :605
 Median :  0.2340   Median :1.00268   Median :  0.2410
 Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499
 3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050
 Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260
```



Volume 和 Year 有指數和對數關係。

(b)

```
> fit <- glm(Direction ~.-Today -Year, data=Weekly, family="binomial")
> summary(fit)

Call:
glm(formula = Direction ~ . - Today - Year, family = "binomial",
    data = Weekly)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6949  -1.2565   0.9913   1.0849   1.4579

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.26686    0.08593   3.106   0.0019 **
Lag1        -0.04127    0.02641  -1.563   0.1181
Lag2         0.05844    0.02686   2.175   0.0296 *
Lag3        -0.01606    0.02666  -0.602   0.5469
Lag4        -0.02779    0.02646  -1.050   0.2937
Lag5        -0.01447    0.02638  -0.549   0.5833
Volume      -0.02274    0.03690  -0.616   0.5377
```

Lag2 在 alpha=0.05 時,此變數有顯著效果

(c)

```
> glm.probs <- predict(fit,type="response")
> class.glm <- car::recode(glm.probs,"0:0.499999999='down';0.5:1='Up'")
> table(class.glm ,Weekly$Direction)

class.glm Down  Up
     down   54  48
     Up    430 557
```

48 和 430 是預測錯誤的次數,48 是預測 down 結果是 up,430 是預測 up 結果是 down

準確率=54+557/(54+48+430+557)=54.511

(D)

```
> train <- (Year<2009)
> test <-  Weekly[!train ,]
> fit1 <- glm(Direction ~ Lag2, data=Weekly, subset=train, family="binomi
al")
> glm.probs <- predict(fit1, type="response", newdata=test)
> class.glm1 <- car::recode(glm.probs,"0:0.499999999='down';0.5:1='Up'")
> table(class.glm1 ,test$Direction)

class.glm1 Down Up
      down    9  5
      Up     34 56
> (9+56)/(9+5+34+56)
[1] 0.625
```

(e)

```
> fit1 <- lda(Direction ~ Lag2, data=Weekly, subset=train, family="binomi
al")
> glm.probs <- predict(fit1, type="response", newdata=test)
> table(glm.probs$class,test$Direction)

       Down Up
  Down    9  5
  Up     34 56
```

(f)

```
> fit1 <- qda(Direction ~ Lag2, data=Weekly, subset=train, family="binomi
al")
> glm.probs <- predict(fit1, type="response", newdata=test)
> table(glm.probs$class,test$Direction)

       Down Up
  Down    0  0
  Up     43 61
```

61/(43+61)=0.5865

(g)

```
> library(class)
> train.k = Weekly[train, c("Lag2", "Direction")]
> knn.pred = knn(train=data.frame(train.k$Lag2), test=data.frame(test$Lag
2), cl=train.k$Direction, k=1)
> hah <- table(test$Direction, knn.pred)
> hah
      knn.pred
       Down Up
  Down   21 22
  Up     30 31
> (21+31)/(21+22+30+31)
[1] 0.5
```

(h) logistic 和 lda 準確率是比較高的

11. In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto` data set.

   (a) Create a binary variable, `mpg01`, that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median. You can compute the median using the `median()` function. Note you may find it helpful to use the `data.frame()` function to create a single data set containing both `mpg01` and the other `Auto` variables.

   (b) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

   (c) Split the data into a training set and a test set.

   (d) Perform LDA on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

   (e) Perform QDA on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

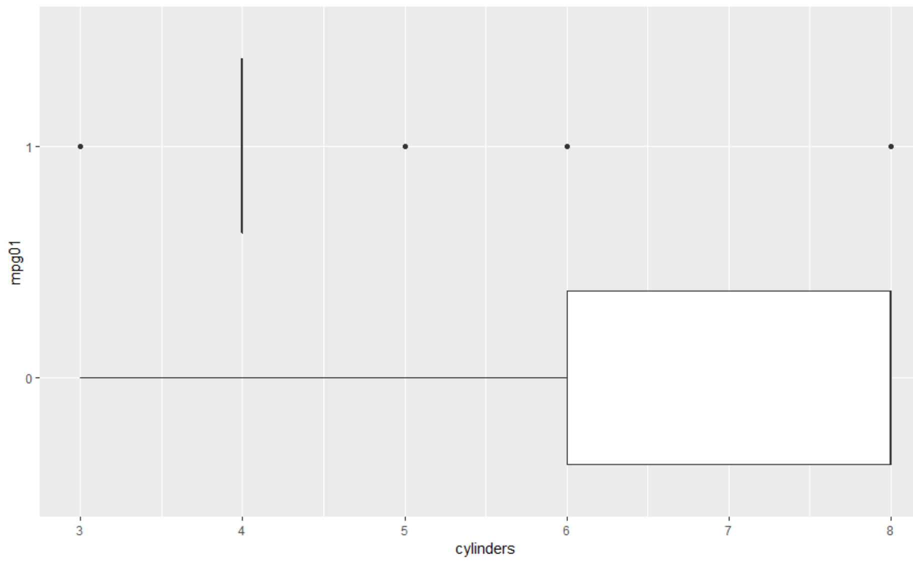
   (f) Perform logistic regression on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

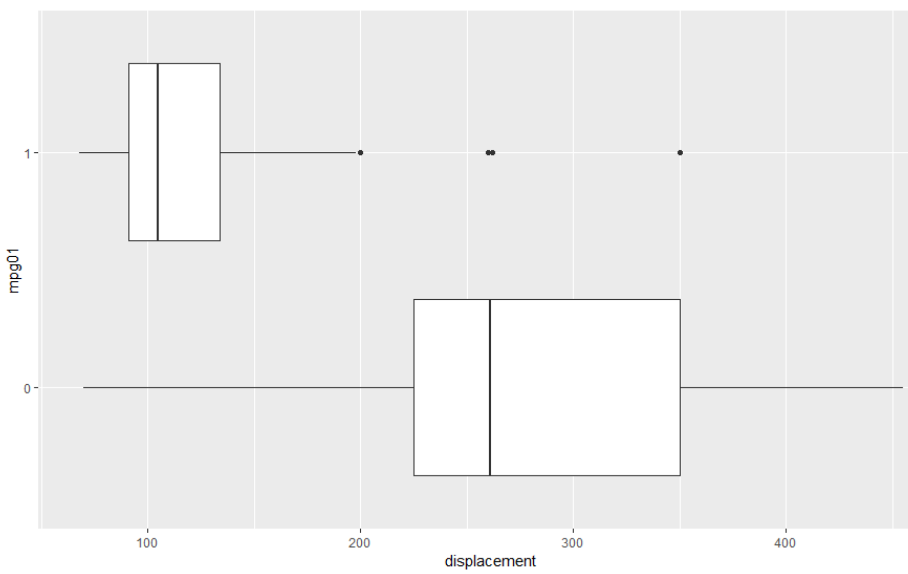
   (g) Perform KNN on the training data, with several values of K, in order to predict `mpg01`. Use only the variables that seemed most associated with `mpg01` in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?

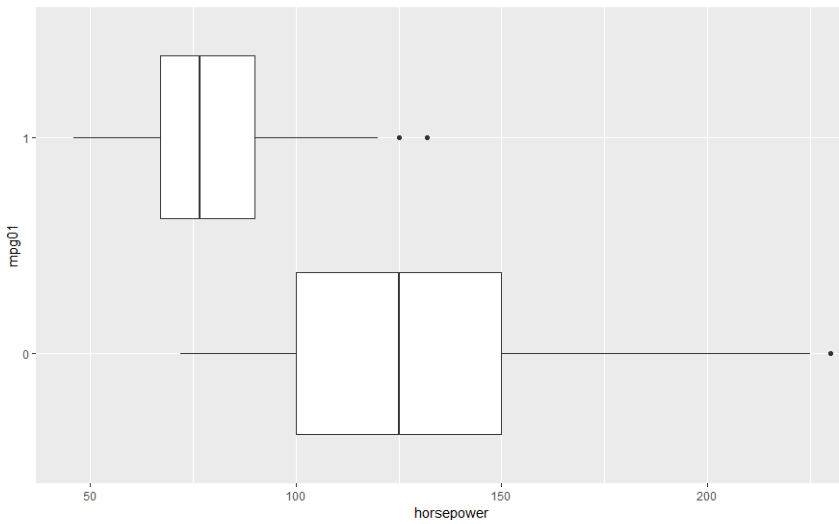
(a)

```r
Auto$mpg01 <- ifelse(Auto$mpg>median(Auto$mpg), "1", "0")
```

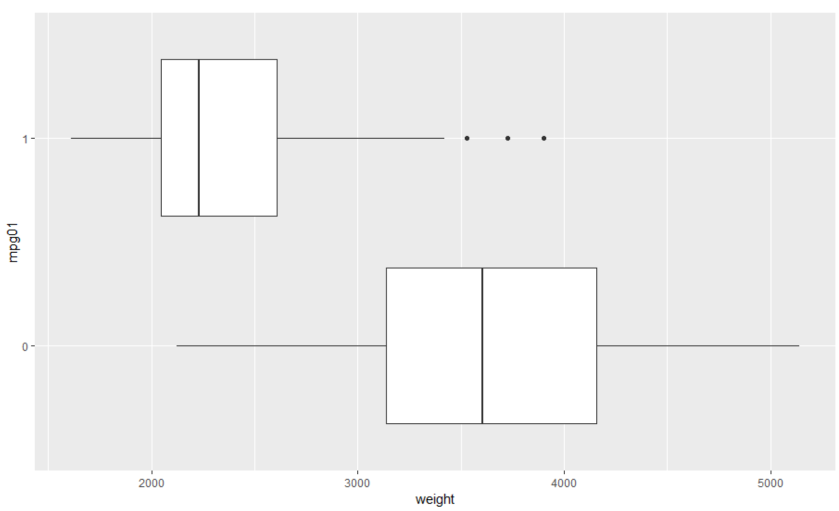
(b)



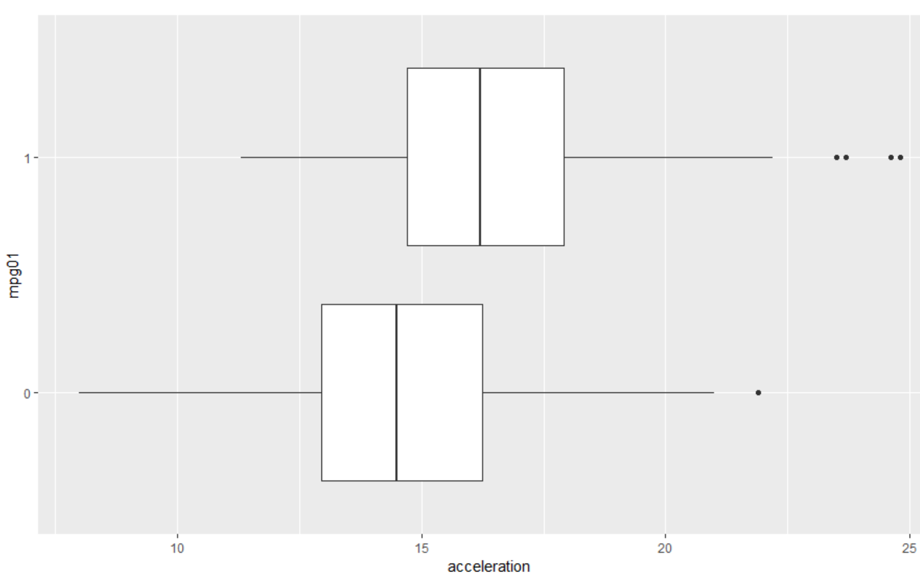




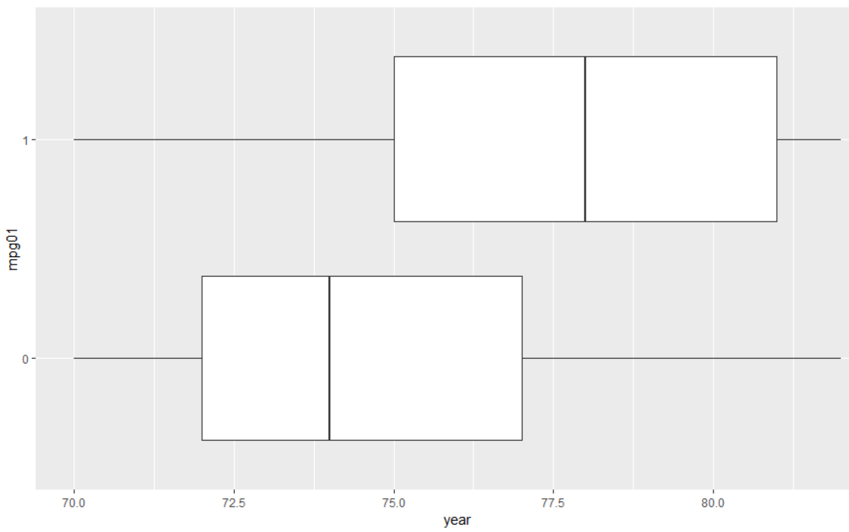
mpg01
1
0
50
100
150
200
horsepower



mpg01
1
0
2000
3000
4000
5000
weight



mpg01
1
0
10
15
20
25
acceleration



(C)

```
rows <- sample(x=nrow(Auto), size=.75*nrow(Auto))
train <- Auto[rows, ]
test <- Auto[-rows, ]
```

(d)

```
> lda.fit <- lda(mpg01 ~ displacement+horsepower+weight+acceleration+year
+cylinders+origin, data=train)
> lda.pred <- predict(lda.fit, test)
> table(test$mpg01, lda.pred$class)

     0  1
  0 40 11
  1  2 45
```

Test error:(11+2)/(40+45+11+2)=0.132

(E)

```
> qda.fit <- qda(mpg01 ~ displacement+horsepower+weight+acceleration+year
+cylinders+origin, data=train)
> qda.pred <- predict(qda.fit, test)
> table(test$mpg01, qda.pred$class)

     0  1
  0 44  7
  1  2 45
```

Test error: 9/(9+44+45)=0.09

(f)

```
> fit1 <- glm(as.factor(mpg01) ~ displacement+horsepower+weight+accelerat
ion+year+cylinders+origin, data=train, family="binomial")
> glm.probs <- predict(fit1, type="response", newdata=test)
> class.glm1 <- car::recode(glm.probs,"0:0.499999999='down';0.5:1='Up'")
> table(class.glm1 ,test$mpg01)

class.glm1  0  1
      down 43  2
      Up    8 45
```

Test error=10/(10+43+45)=0.0102