Assignment 2
Instruction:

- You can use this WORD file as an answer sheet. Attach required output below each question. R codes can be attached to each question or at the end of the document (for partial credits if your results are not correct).
- Name the word file as: YourName.doc

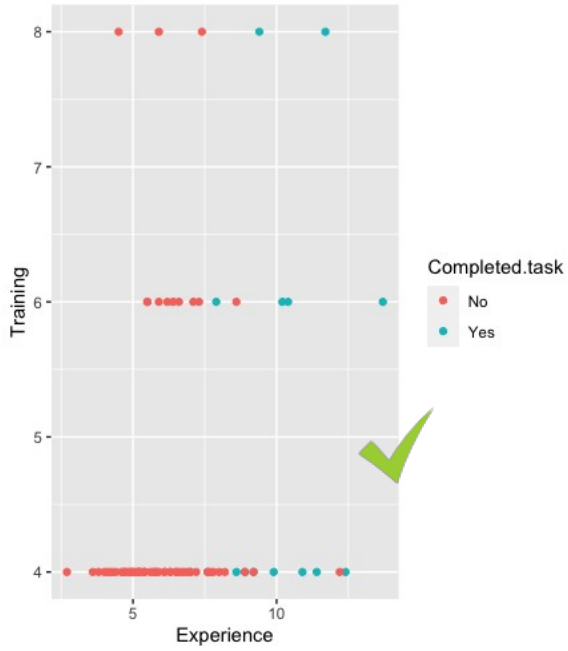
A management consultant is studying the roles played by experience and training in a system administrator's ability to complete a set of tasks in a specified amount of time. In particular, she is interested in discriminating between administrators who are able to complete given tasks within a specified time and those who are not. Data are collected on the performance of 75 randomly selected administrators. They are stored in the file *SystemAdministrators.csv*.

The variable Experience measures months of full-time system administrator experience, while Training measures the number of relevant training credits. The outcome variable Completed Task is either Yes or No, according to whether or not the administrator completed the tasks.

Part 1: Exploratory Analysis

(1) Create a scatter plot of Experience versus Training using color or symbol to distinguish programmers who completed the task from those who did not complete it. Which predictor(s) appear(s) potentially useful for classifying task completion?

Attach scatter plot here.



Which predictor(s) appear(s) potentially useful for classifying task completion?
Answer: Experience

Part 2: Logistic Regression

(2) Partition 80% of the data into a training set, and 20% into a validation set. When you sample for training data, run **set.seed(2)** for reproducible samples. Run a logistic regression model to predict probability of completed task with both predictors.

Attach the model summary here.

```
Call:
glm(formula = Completed ~ Experience + Training, family = "binomial",
    data = train.df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.54597  -0.33529  -0.15534  -0.06406   2.30280

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.5585     3.2776  -3.527 0.000421 ***
Experience    1.1114     0.3107   3.577 0.000347 ***
Training      0.3001     0.3589   0.836 0.403108
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62.719  on 59  degrees of freedom
Residual deviance: 27.397  on 57  degrees of freedom
AIC: 33.397

Number of Fisher Scoring iterations: 6
```

(3) Write the estimated logistic regression model.
Log(p/1-p) =-11.5585+Experience*1.1114+Traning*0.3001#log odds
odds = exp(log.odds)
# probability of acceptance: odds/(1+odds)

(4) Interpret the effect of Experience on Completed Task.
Log odds of completion increases by 1.1114 as experience increases by 1, when other variables are held constant.

(5) Use AIC based backward elimination to choose predictors.

Attach the summary of the model picked by backward elimination here.

```
> summary(backward)

Call:
glm(formula = Completed ~ Experience, family = "binomial", data = train.df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.66644  -0.34944  -0.16572  -0.07161   2.19548

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -10.180      2.617  -3.889 0.000101 ***
Experience     1.123      0.309   3.636 0.000277 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62.719  on 59  degrees of freedom
Residual deviance: 28.093  on 58  degrees of freedom
AIC: 32.093

Number of Fisher Scoring iterations: 6
```

(6) An administrator has 6 training credits and 7 months’ experience. Use model in (5) to predict the administrator’s probability of task completion. What is the odds of task completion for this administrator?
Log(p/1-p) = -10.180 + 1.123 * 7 = -1.681 4.1
Odds=e^(-1.681)= 0.186

(7) Apply the model in part (5) to the validation data set to predict probability of task completion. Use 0.5 as a threshold for classifying the predicted probabilities as either 1 (task completed) or 0 (task not completed). Construct a confusion matrix for the predictions.

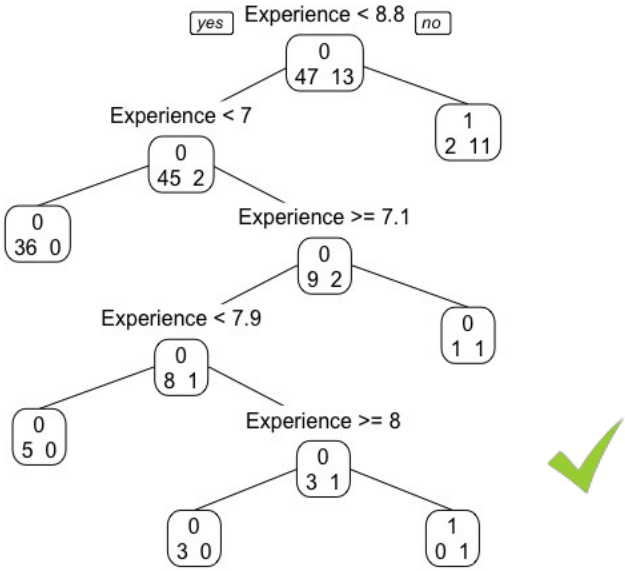
```
> conf.mat1
          Actual
Prediction  0  1
         0 13  2
```

(8) Calculate the error rate of predictions in (7).
(2/15)*100%=13%

Part 3: Classification Tree

(9) Use a classification tree to predict completed task with the two predictors, using the **training set**. Grow a large tree using the options: minbucket=1, cp=0.0001. Plot the classification tree.

Attach the tree plot here.



(10) For the root node (top node of the tree plot), calculate its Gini index.

0.3[illegible]

(11) Which class will an administrator with Experience=8, and Training=6 be classified to based on the tree in part (9)?

0

(12) Do you think such a large tree in (9) will have good prediction accuracy for unseen data?

No, it uses 'Experience' multiple times, it creates too much noises.

(13) Use 10-fold cross validation to determine the best value of complexity parameter (it controls the tree size). **Before running cross validation, run set.seed(3)**. Use the options **minbucket=1, cp=0.0001** for cross validation for the biggest tree.

Attach the cp table here.

```
Variables actually used in tree construction:
[1] Experience

Root node error: 13/60 = 0.21667

n= 60

        CP nsplit rel error  xerror    xstd
1 0.692308      0   1.00000 1.00000 0.24547
2 0.019231      1   0.30769 0.30769 0.14863
3 0.000100      5   0.23077 0.61538 0.20255
>
```

(14) What complexity parameter (cp) value minimizes the cross validation error in (13)?

0.019231

Index of comments