Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name) 1. Bian, Yu 2. Huang, Ruiqi 3. Liu, Ching-Wei	4. Liu, Mengjie 5. Wang, Zhaofei 6. 7. **GROUP NUMBER:** **5**
MSc in: Business Analytics	
Module Code: SMM634	
Module Title: Analytics Methods for Business	
Lecturer: Rosalba Radice	**Submission Date:** 19/03/2019

Declaration:

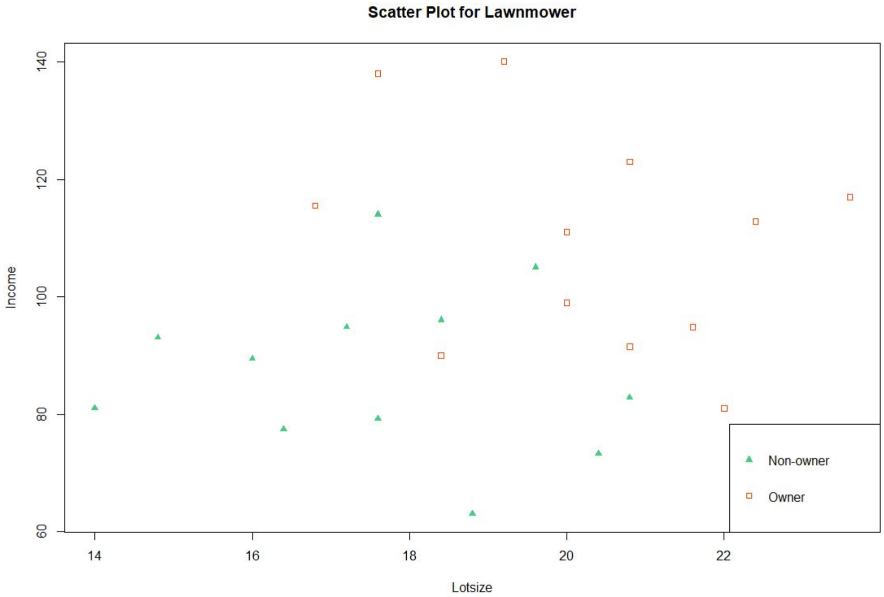
By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission: | **Final Mark:** %

Question 1



There's a rough linear boundary between owners and non-owners, therefore the two variables should be sufficient in classifying the dataset.

Question 2

```
Call:
lda(owner ~ ., data = lawnmower)

Prior probabilities of groups:
nonowner    owner
     0.5      0.5

Group means:
           income  lotsize
nonowner   87.400 17.63333
owner     109.475 20.26667

Coefficients of linear discriminants:
              LD1
income  0.0484468
lotsize 0.3795228
```

confusion matrix:

```
          nonowner owner
 nonowner       10     1
 owner           2    11
```

Accuracy rate:

```
[1] 0.875
```

Prior probability is 0.5 for both classes, meaning the original dataset contains equal number of owners and non-owners. The group mean indicates on average owners have a high income and lotsize than non-owners. The coefficient is simply used to estimate the probability that the observation is an owner (this can be clarified using contrasts() function in R), therefore the higher the value of 0.048*income+0.38*lotsize, the higher probability the observation is an owner, vice versa. We use the same dataset to predict the owner status and result in a 0.875 accuracy in predicting the result, which is given by both confusion matrix and the mean() function. Therefore, we could say that the training error is 1-0.875=12.5%, as we are using the training dataset to predict. The error in this case might be underestimated, and we should separate the original dataset into training and test subset in order to obtain a more unbiased test error.

Question 3

```
Call:
lda(owner ~ ., data = lawnmower, subset = train.index)

Prior probabilities of groups:
nonowner    owner
     0.5      0.5

Group means:
          income lotsize
nonowner   85.20   17.32
owner     110.79   20.16

Coefficients of linear discriminants:
               LD1
income  0.04906178
lotsize 0.35609743
```

confusion matrix:

```
         nonowner owner
nonowner        1     0
owner           1     2
```

Accuracy rate:

```
[1] 0.75
```

As we expected, the test error now is generated to be 1-0.75=0.25, which is higher than before. The result is more valid than before as we now using different dataset to test the model prediction power. However, the original dataset is still too small that the random sampling bias still massively impact the result. For example, if we change the seed to 100, the test error would turn out to be 50%. We need a bigger dataset to mitigate the sampling bias.

Question 4

```
Call:
qda(owner ~ ., data = lawnmower, subset = train.index)

Prior probabilities of groups:
nonowner    owner
     0.5      0.5

Group means:
          income lotsize
nonowner   85.20   17.32
owner     110.79   20.16
```

confusion matrix:

```
         nonowner owner
nonowner        1     0
owner           1     2
```

Accuracy rate:

```
[1] 0.75
```

QDA model displays the same prediction accuracy as LDA with the same test error of 0.75. This might be because the dataset is too small so that the two methods hardly differ from each other. Roughly speaking, LDA tends to be a better bet than QDA if there are relatively few training observations and so reducing variance is crucial. In contrast, QDA is recommended if the training set is very large, so that the variance of the classifier is not a major concern, or if the assumption of a common covariance matrix for the K classes is clearly untenable.

Question 5

We assume that the data set for this question is called “default”, the response variable in this data set is “jobclassifications”, and the three predictors are “outdooractivity”, “sociability” and “conservativeness”. Therefore, we could fit the LDA model as follows:

```
# divide the original data into training and test dataset
set.seed(328)
train_index <- sample(seq_len(nrow(default)), size = 50)# random sampling 50 observation as training
train.default <- default[train_index, ]
test.default <- default[-train_index, ]
# fit lda into the training dataset
lda.fit.default <- lda(jobclassification~outdooractivity+sociability+conservativeness, data = train.default)
# use lda model to predict test dataset
lda.pred.default <- predict(lda.fit.default, test.default)
```