DS502- HW2

Mahdi Alouane and Rahul Pande

1.(5 points) Section 4.7, page 168, question 1

2.(10 points) Section 4.7, page 169-170, question 5

3.(5 points) Section 4.7, page 170, question 8

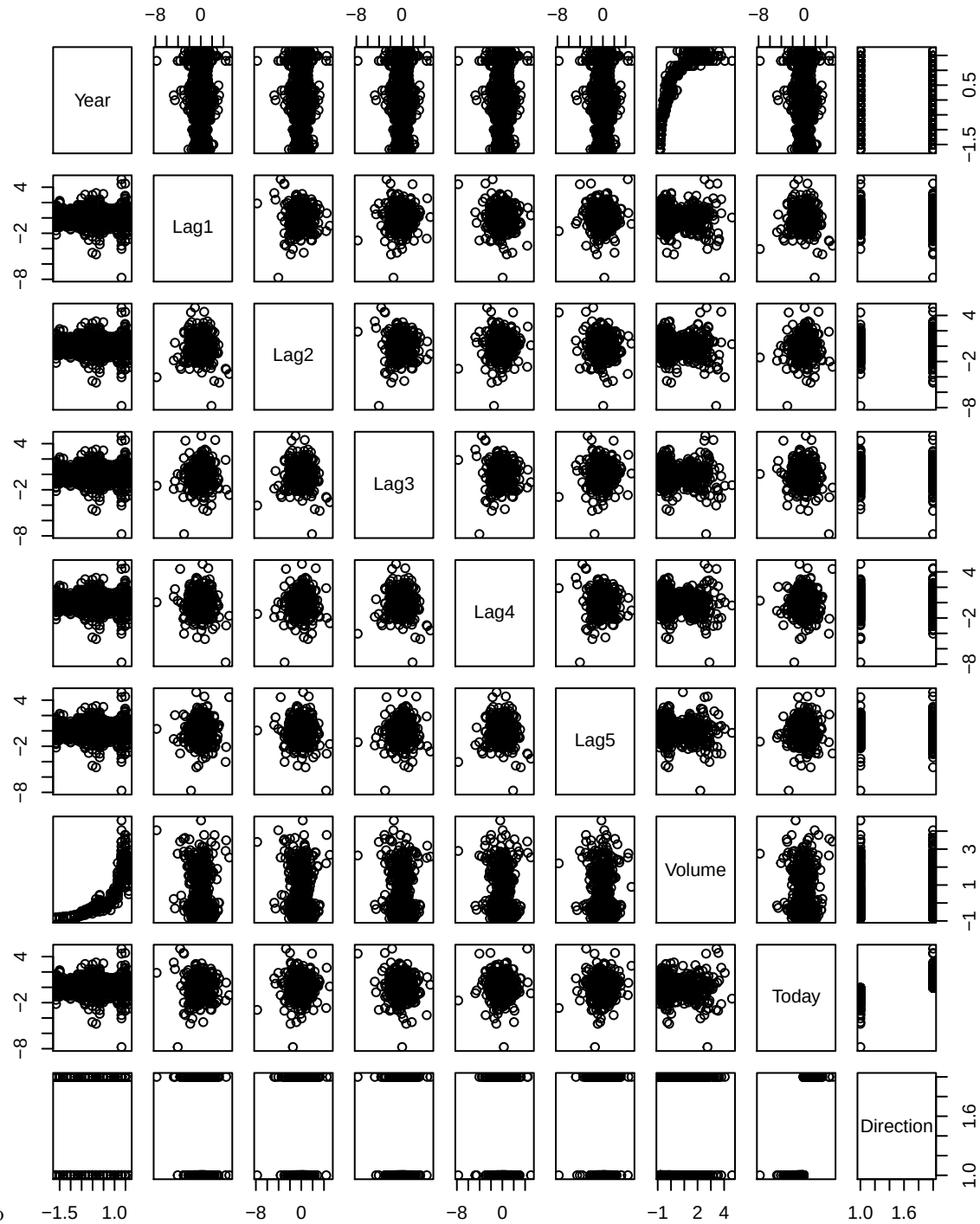
4.(15 points) Section 4.7, page171, question 10

a.

```
require(ISLR)
```

```
## Loading required package: ISLR
```

```
data("Weekly")
data.weekly = Weekly
summary(data.weekly)

# standardize data
data.weekly[ , colnames(data.weekly) != "Direction"] = scale(data.weekly[ , colnames(data.weekly) != "Di

# pair plot
pairs(data.weekly)
```

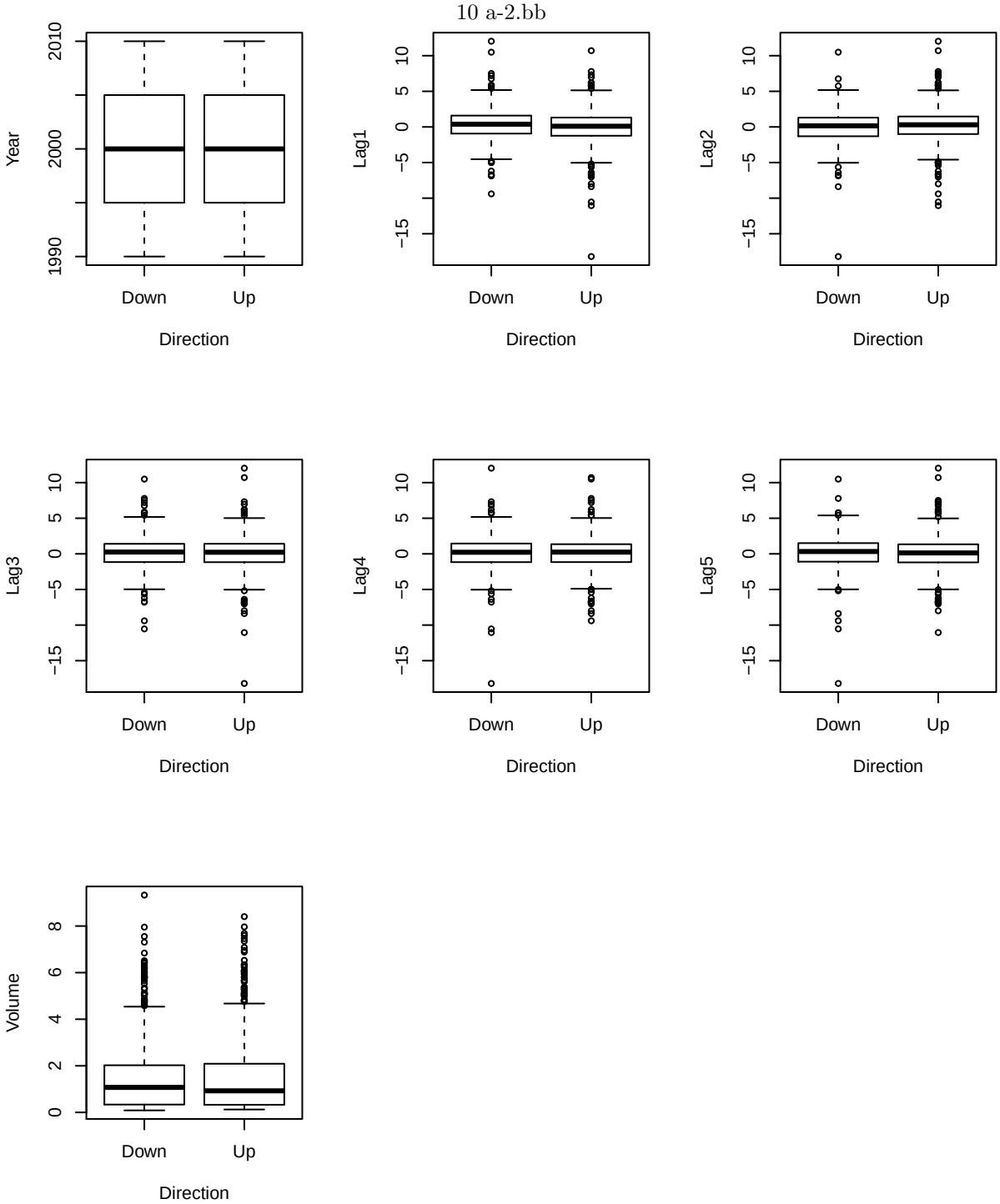


10 a-1.bb

```
# box plots w.r.t. response variables
par(mfrow=c(3,3))

vars = setdiff(names(Weekly), c("Today", "Direction"))

plot_against_direction <- function(y){
  x = "Direction"
  f <- as.formula(paste(c(y, x), collapse = " ~ "))
  boxplot(f, data= Weekly, ylab = y, xlab = x)
```

```r
}

sapply(vars, plot_against_direction)
```

10 a-2.bb



- `Year` is non-linearly related to `Volume`. At first, `Volume` increases rapidly with `Year` but then becomes constant.
- From the shape of graphs of `Year` against `Lag` variables, the variability in `Lag` vars is higher in the middle and in the end `Year`

- Similarly, the variablility in `Lag` variables is higher towards either end (high or low) of `Volume` than the middle region.

b.

```
logistic.fit <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = data.weekly, family =
summary(logistic.fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = "binomial", data = 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)  0.22523    0.06128   3.676 0.000237 ***
## Lag1        -0.09727    0.06225  -1.563 0.118144
## Lag2         0.13776    0.06333   2.175 0.029601 *
## Lag3        -0.03791    0.06294  -0.602 0.546924
## Lag4        -0.06559    0.06246  -1.050 0.293653
## Lag5        -0.03417    0.06230  -0.549 0.583348
## Volume      -0.03836    0.06223  -0.616 0.537675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

From the `summary` of the fit, we see that `Lag2` is a significant factor in predicting `Direction`

c.

```
pred <- round(predict(logistic.fit, data.weekly, type = "response"))
pred <- factor(x = pred, labels = c("Down", "Up"))
actual <- data.weekly$Direction
conf_matrix <- table(actual, pred)
message("Confusion matrix")
```

```
## Confusion matrix
```

```
conf_matrix
```

```
##       pred
## actual Down  Up
##   Down   54 430
##   Up     48 557
```

```
message("Fraction of correct predictions:")
```

```
## Fraction of correct predictions:
```

```r
(conf_matrix[1] + conf_matrix[4])/sum(conf_matrix)
```

```
## [1] 0.5610652
```

Consider that the logistic model is predicting "is the return positive (`Up`)?" Now from the confusion matrix the model is making `430` **type 1** errors and `48` **type 2** errors

d.

```r
train.data <- data.weekly[Weekly$Year <= 2008, ]
heldout.data <- data.weekly[Weekly$Year > 2008, ]

lag2.fit <- glm(Direction ~ Lag2, data = train.data, family = "binomial")

lag2_pred <- round(predict(lag2.fit, heldout.data, type = "response"))
lag2_pred <- factor(x = lag2_pred, labels = c("Down", "Up"))
lag2_actual <- heldout.data$Direction
lag2_conf_matrix <- table(lag2_actual, lag2_pred)
message("Confusion matrix for Lag2 model")
```

```
## Confusion matrix for Lag2 model
```

```r
lag2_conf_matrix
```

```
##            lag2_pred
## lag2_actual Down Up
##        Down    9 34
##        Up      5 56
```

```r
message("Fraction of correct predictions:")
```

```
## Fraction of correct predictions:
```

```r
(lag2_conf_matrix[1] + lag2_conf_matrix[4])/sum(lag2_conf_matrix)
```

```
## [1] 0.625
```

e.

```r
require(MASS)
```

```
## Loading required package: MASS
```

```r
lda.fit <- lda(Direction ~ Lag2, data = train.data)

lda_pred <- predict(lda.fit, heldout.data)$class
lda_actual <- heldout.data$Direction
lda_conf_matrix <- table(lda_actual, lda_pred)
message("Confusion matrix for lda model")
```

```
## Confusion matrix for lda model
```

```r
lda_conf_matrix
```

```
##           lda_pred
## lda_actual Down Up
##       Down    9 34
##       Up      5 56
```

```r
message("Fraction of correct predictions:")
```

```
## Fraction of correct predictions:
```

```r
(lda_conf_matrix[1] + lda_conf_matrix[4])/sum(lda_conf_matrix)
```

```
## [1] 0.625
```

f.

```r
qda.fit <- qda(Direction ~ Lag2, data = train.data)

qda_pred <- predict(qda.fit, heldout.data)$class
qda_actual <- heldout.data$Direction
qda_conf_matrix <- table(qda_actual, qda_pred)
message("Confusion matrix for lda model")
```

```
## Confusion matrix for lda model
```

```r
qda_conf_matrix
```

```
##           qda_pred
## qda_actual Down Up
##       Down    0 43
##       Up      0 61
```

```r
message("Fraction of correct predictions:")
```

```
## Fraction of correct predictions:
```

```r
(qda_conf_matrix[1] + qda_conf_matrix[4])/sum(qda_conf_matrix)
```

```
## [1] 0.5865385
```

g.

```r
require(class)
```

```
## Loading required package: class
```

```r
set.seed(123)

trainX <- as.matrix(train.data$Lag2)
testX <- as.matrix(heldout.data$Lag2)
knn_pred <- knn(trainX, testX, train.data$Direction, k=1)
knn_actual <- heldout.data$Direction
knn_conf_matrix <- table(knn_actual, knn_pred)
message("Confusion matrix for knn=1 model")
```

```
## Confusion matrix for knn=1 model
```

```r
knn_conf_matrix
```

```
##           knn_pred
## knn_actual Down Up
##       Down   21 22
##       Up     29 32
```

```r
message("Fraction of correct predictions:")
```

```
## Fraction of correct predictions:
```

```r
(knn_conf_matrix[1] + knn_conf_matrix[4])/sum(knn_conf_matrix)
```

```
## [1] 0.5096154
```

h. (i) After experimenting with interaction variables, k and variable transformation, we conclude that Logistic Regression and LDA work best for the `Weekly` dataset.

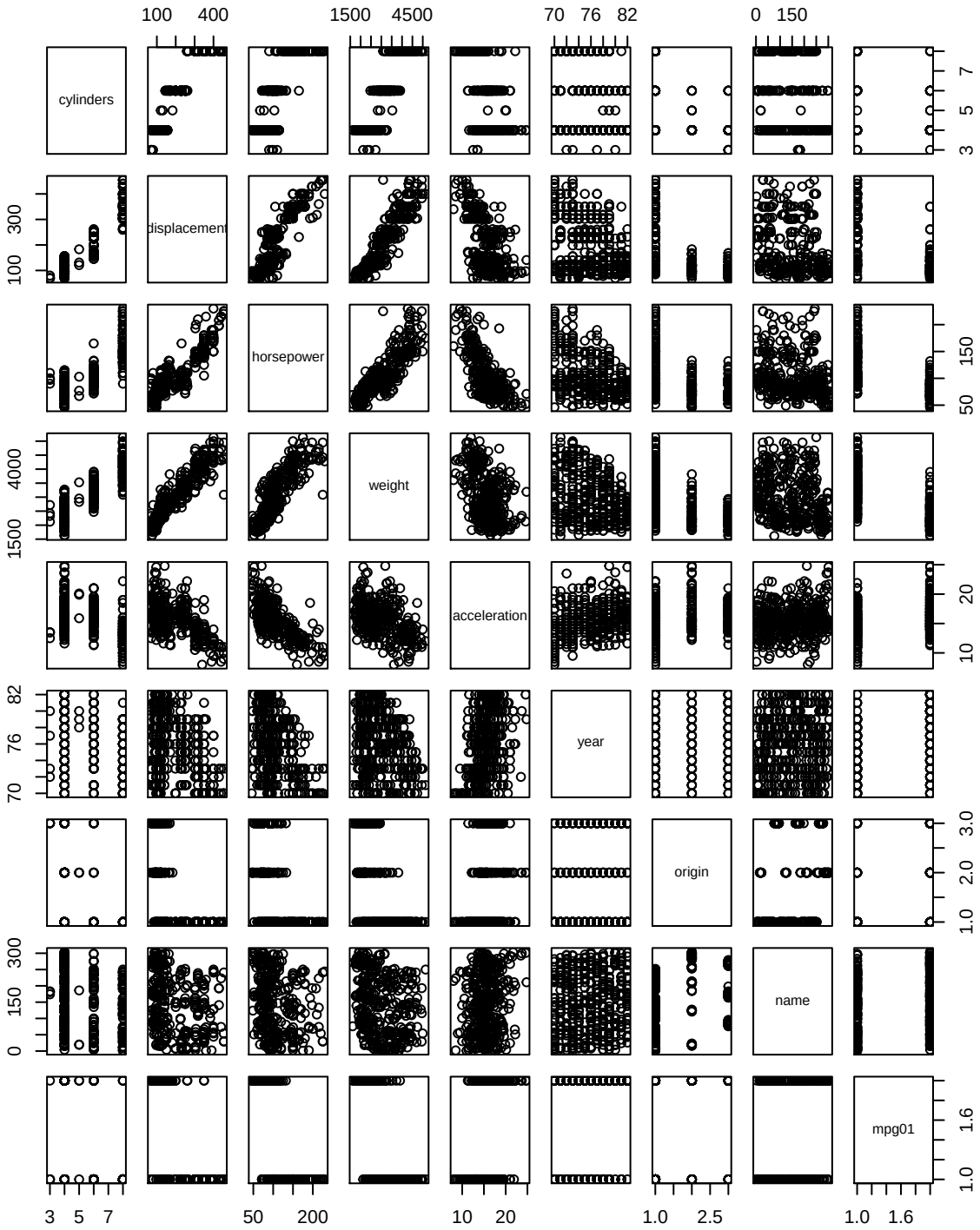
5.(15 points) Section 4.7, page 171-172, question 11

a.

```r
data("Auto")
Auto$mpg01 <- ifelse(Auto$mpg > median(Auto$mpg), 1, 0)
Auto = Auto[, names(Auto) != "mpg"]
Auto$mpg01 <- as.factor(Auto$mpg01)
```

b.

```r
pairs(Auto)
```

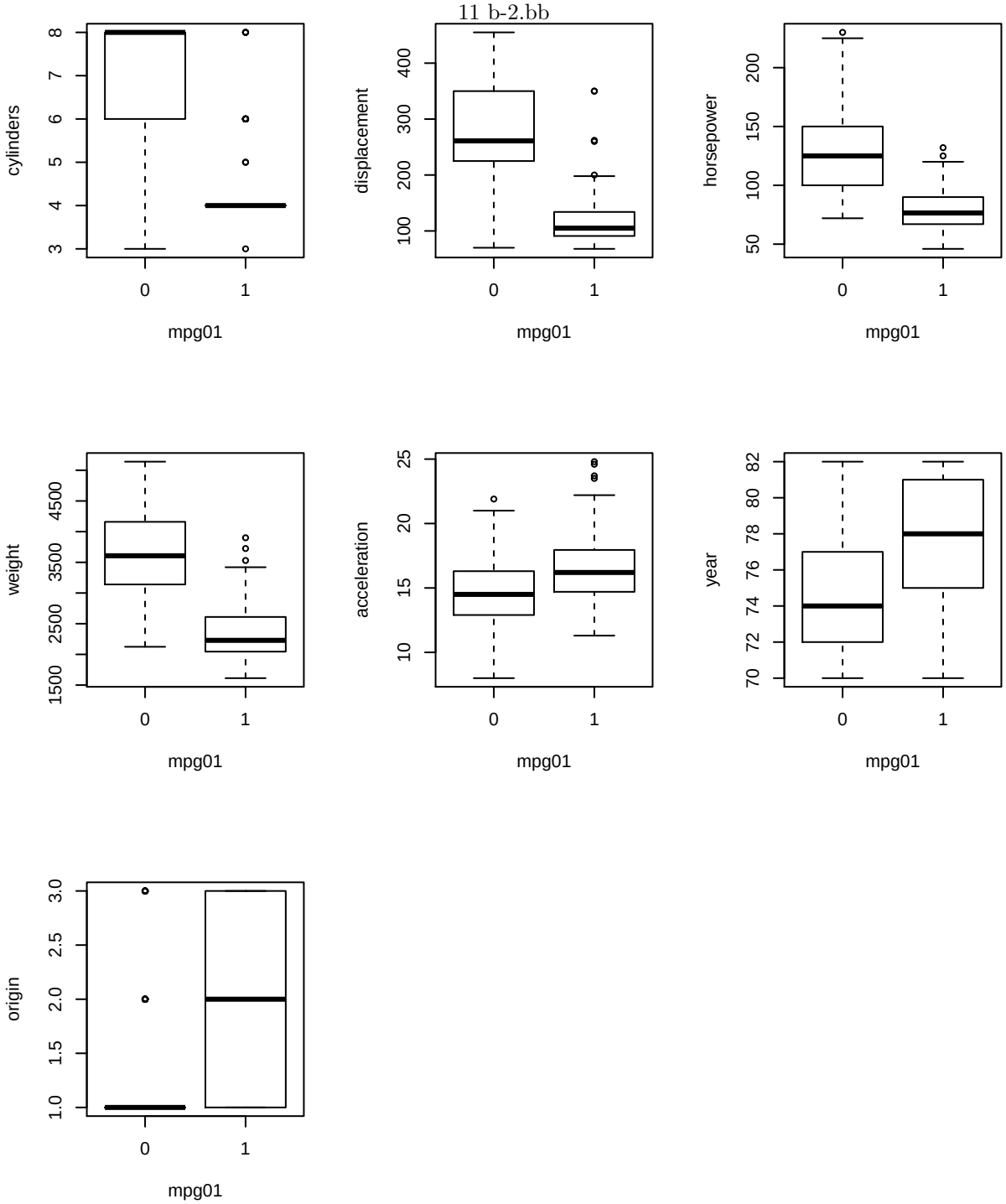


11 b-1.bb

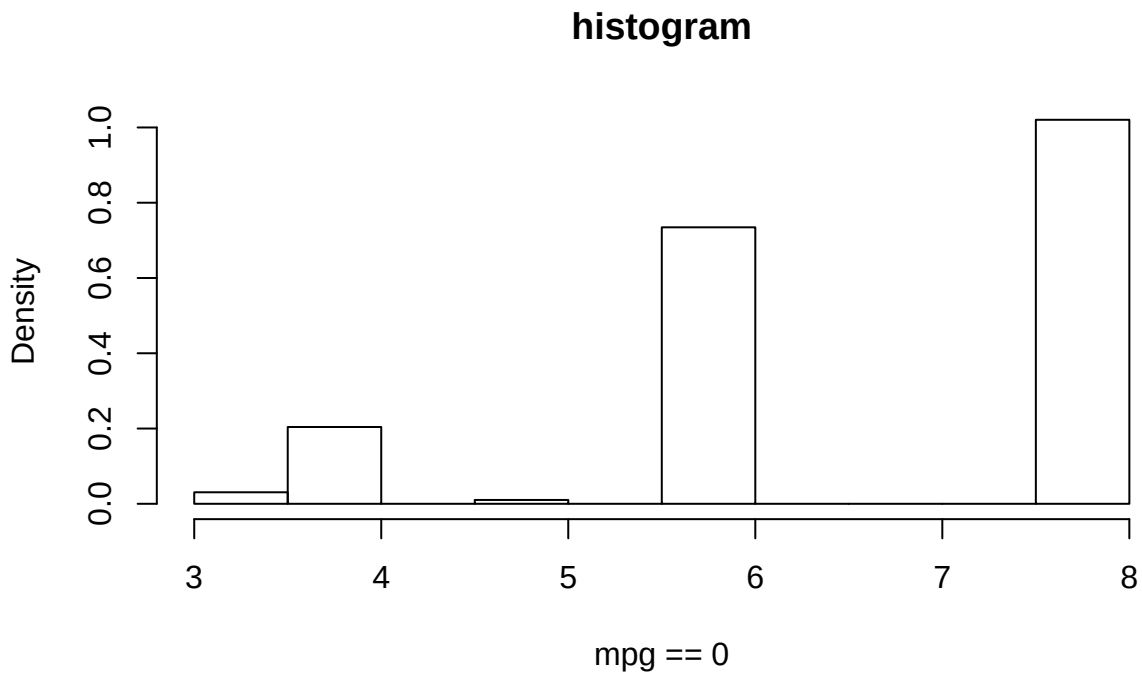
```
par(mfrow=c(3,3))

vars = setdiff(names(Auto), c("mpg01", "name"))

plot_against_mpg01 <- function(y){
  x = "mpg01"
  f <- as.formula(paste(c(y, x), collapse = " ~ "))
  boxplot(f, data = Auto, ylab = y, xlab = x)
}
```

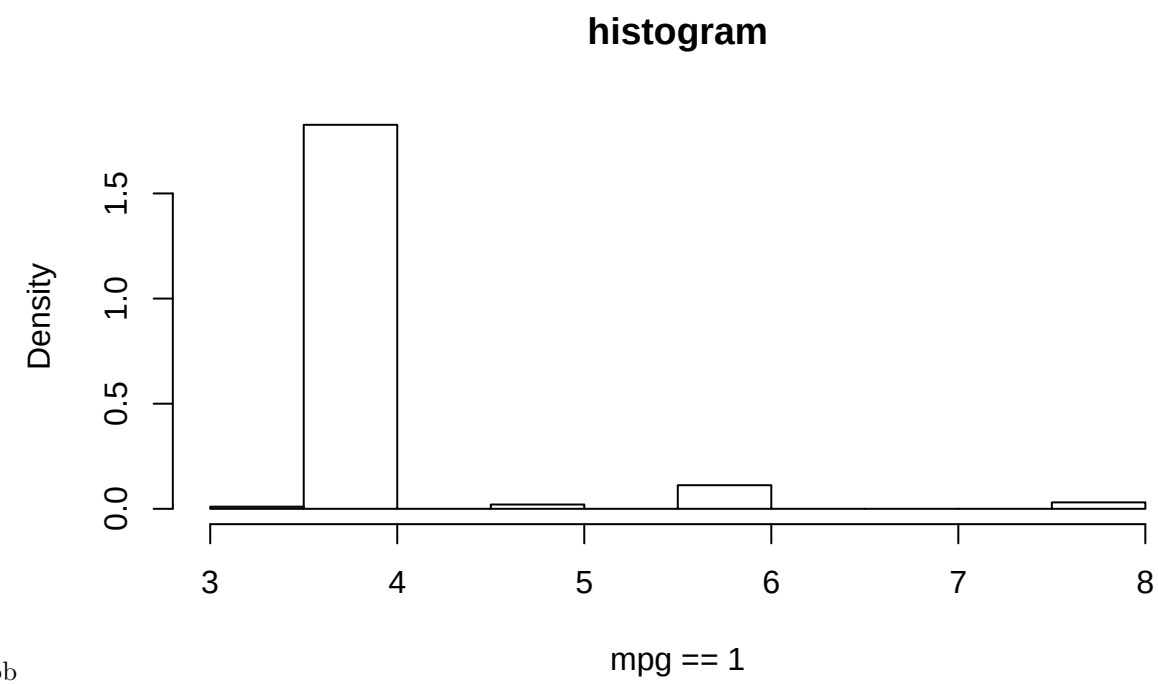
```r
sapply(vars, plot_against_mpg01)

par(mfrow=c(2, 1))
```

11 b-2.bb




```r
hist(Auto$cylinders[Auto$mpg01 == 0 ], xlab = "mpg == 0", freq = F, main = "histogram")
hist(Auto$cylinders[Auto$mpg01 == 1 ], xlab = "mpg == 1", freq = F, main = "histogram")
```





11 b-3.bb

From the plots:

- From the boxplot of `cylinders` vs `mpg01`, `cylinders` is probably the most important predictor of `mpg`. Most of the `1` class are at 4 cylinders.
- `weight` could be a good predictor of `mpg01` since the boxplot shows that the medians are quite apart and overlap is small. Similarly, `displacement` and `horsepower` could also be good predictors.
- There is strong correlation between variables like `weight`, `horsepower`, `displacement`. So there's probably redundant information in these variables.

c.

```
### 6.(10 points) Section 5.4, page 197, question 1
### 7.(10 points) Section 5.4, page 197, question 2
### 8.(15 points) Section 5.4, page 198, question 5
### 9.(15 points) Section 5.4, page 199, question 6
```