Data Science 1 HW1

Vyas Ramankulangara

1592440

In [45]:
```
install.packages("MASS")
library(MASS)
install.packages("RColorBrewer")
library(RColorBrewer)
install.packages("KernSmooth")
library(KernSmooth)
install.packages("caTools")
library(caTools)
install.packages("dplyr")
library(dplyr)
install.packages("rpart")
library(rpart)
install.packages("moments")
library(moments)
```

Number 1

In [2]:
```
# Importing data
dataset <- read.table(file = "data.dat", header = FALSE)
selected_data <- dataset[c(1,2,18)]
```

In [3]:
```
#Calculate covariance matrix
c <- cov(selected_data)
c
correlation <- cor(selected_data)
correlation
```

	V1	V2	V18
V1	67.80657	35.201637	22.391727
V2	35.20164	38.067242	1.775135
V18	22.39173	1.775135	55.335707

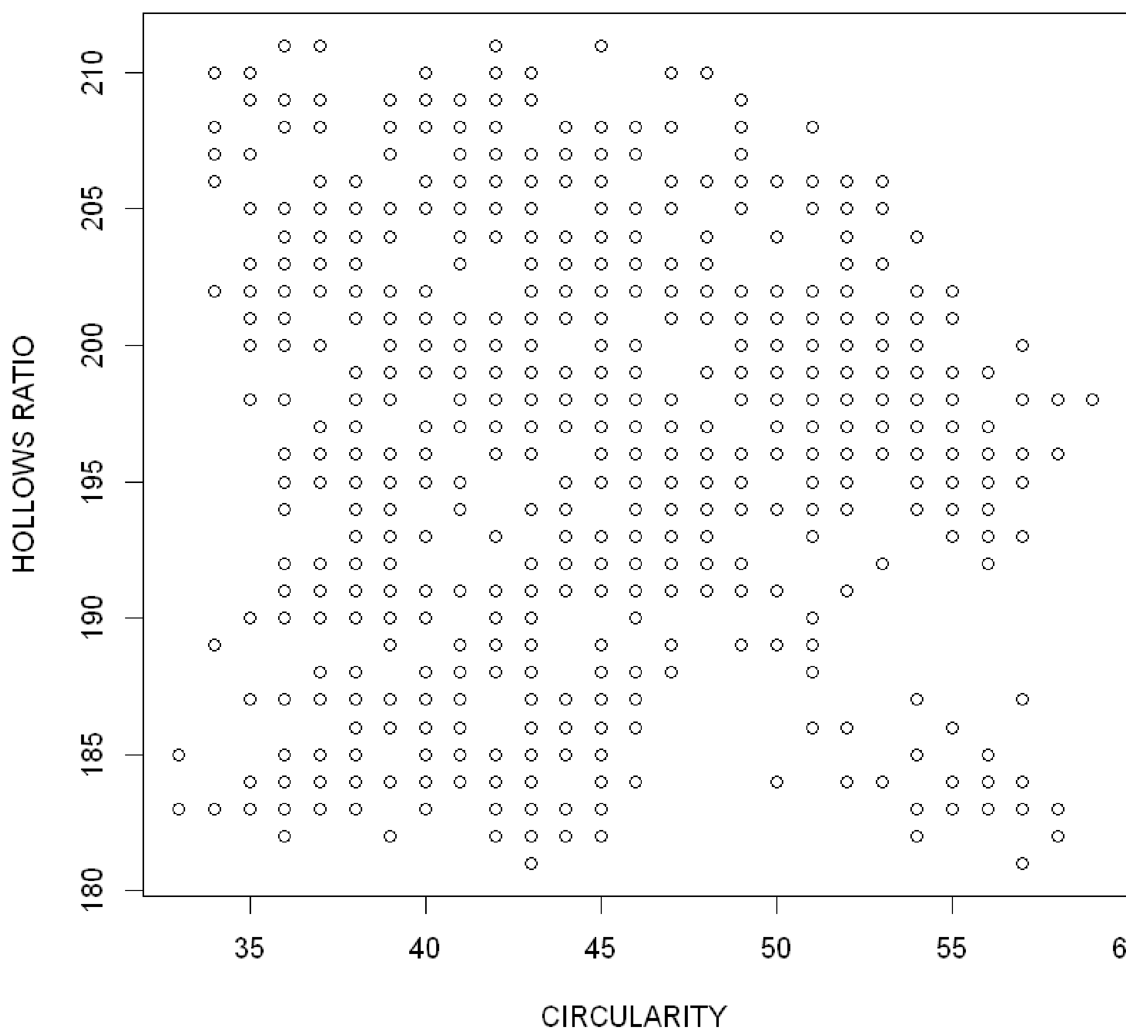
	V1	V2	V18
V1	1.0000000	0.69286923	0.36555185
V2	0.6928692	1.00000000	0.03867702
V18	0.3655518	0.03867702	1.00000000

Analysis:

Correlation is nothing but covariance divided by standard deviation.In the correlation matrix, we can see the relationship between the attributes V1 and V1 is 1 since they are the same attribute and they are related. Any number closer to 1 shows high relation to each other for instance V1 and V2 show high relation.

Number 2

In [4]:
```
# Scatter plot
plot(dataset$V2,dataset$V18, xlab = "CIRCULARITY", ylab = "HOLLOWS RATIO")
```



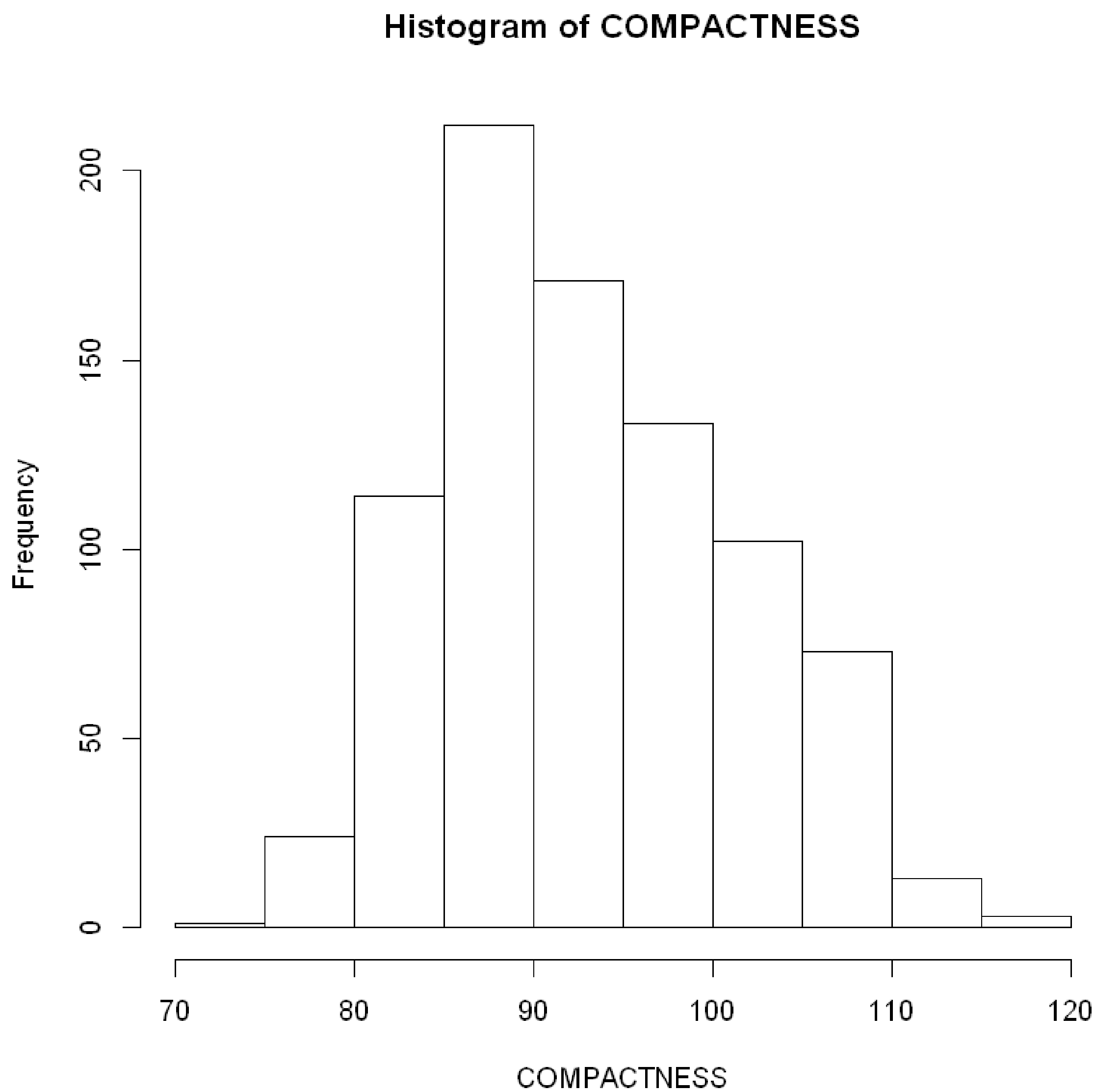
Analysis:

There are no correlation between CIRCULARITY and HOLLOWS RATIO. By observing the graph the values are spread out.

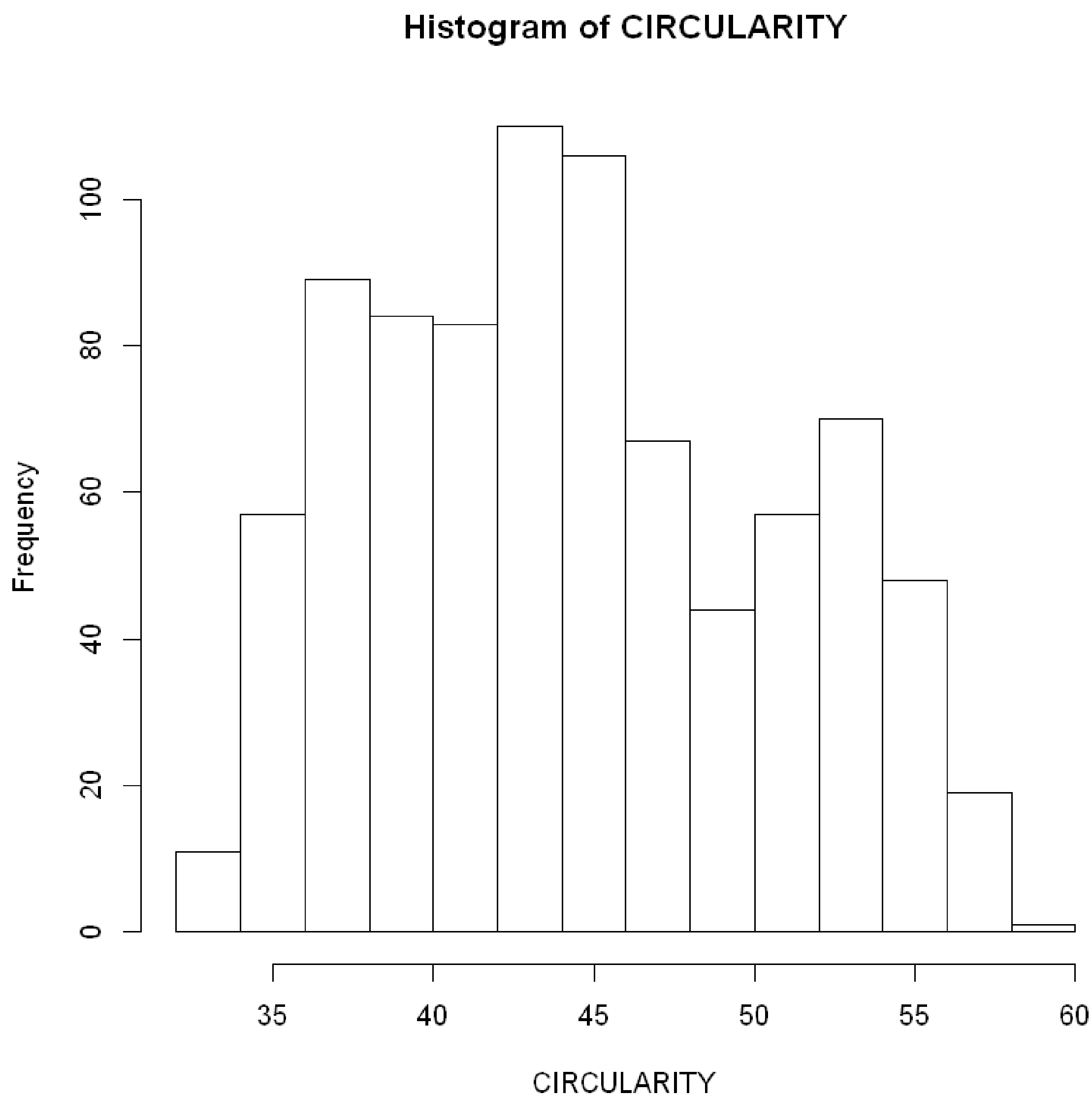
Number 3

In [5]:
```
# Histogram
hist(dataset$V1, main = "Histogram of COMPACTNESS", xlab = "COMPACTNESS", ylab
= "Frequency")
skewness(dataset$V1)
hist(dataset$V2, main = "Histogram of CIRCULARITY", xlab = "CIRCULARITY", ylab
= "Frequency")
skewness(dataset$V2)
```

0.380594287659102



0.262332593782417

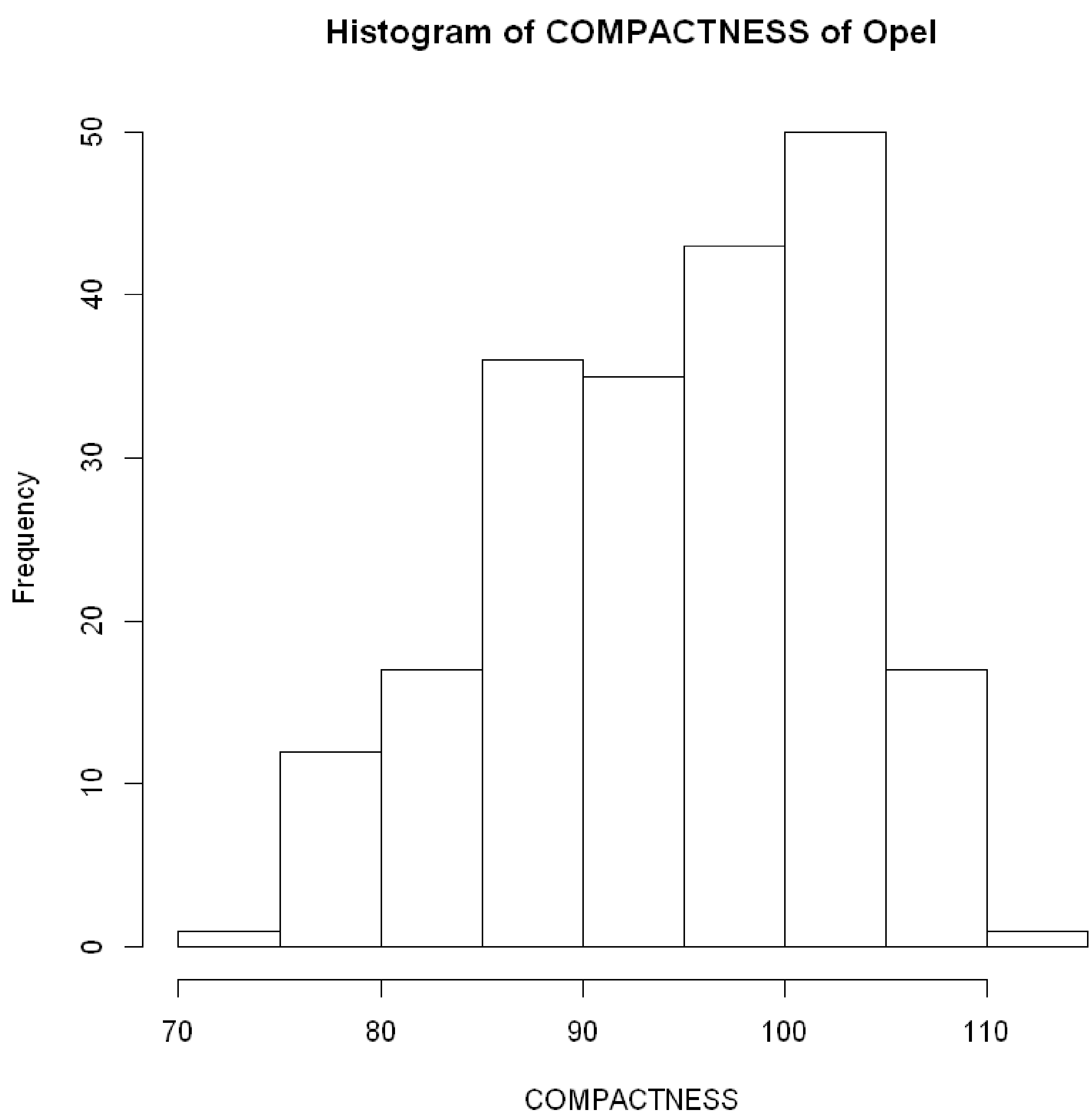


Analysis:

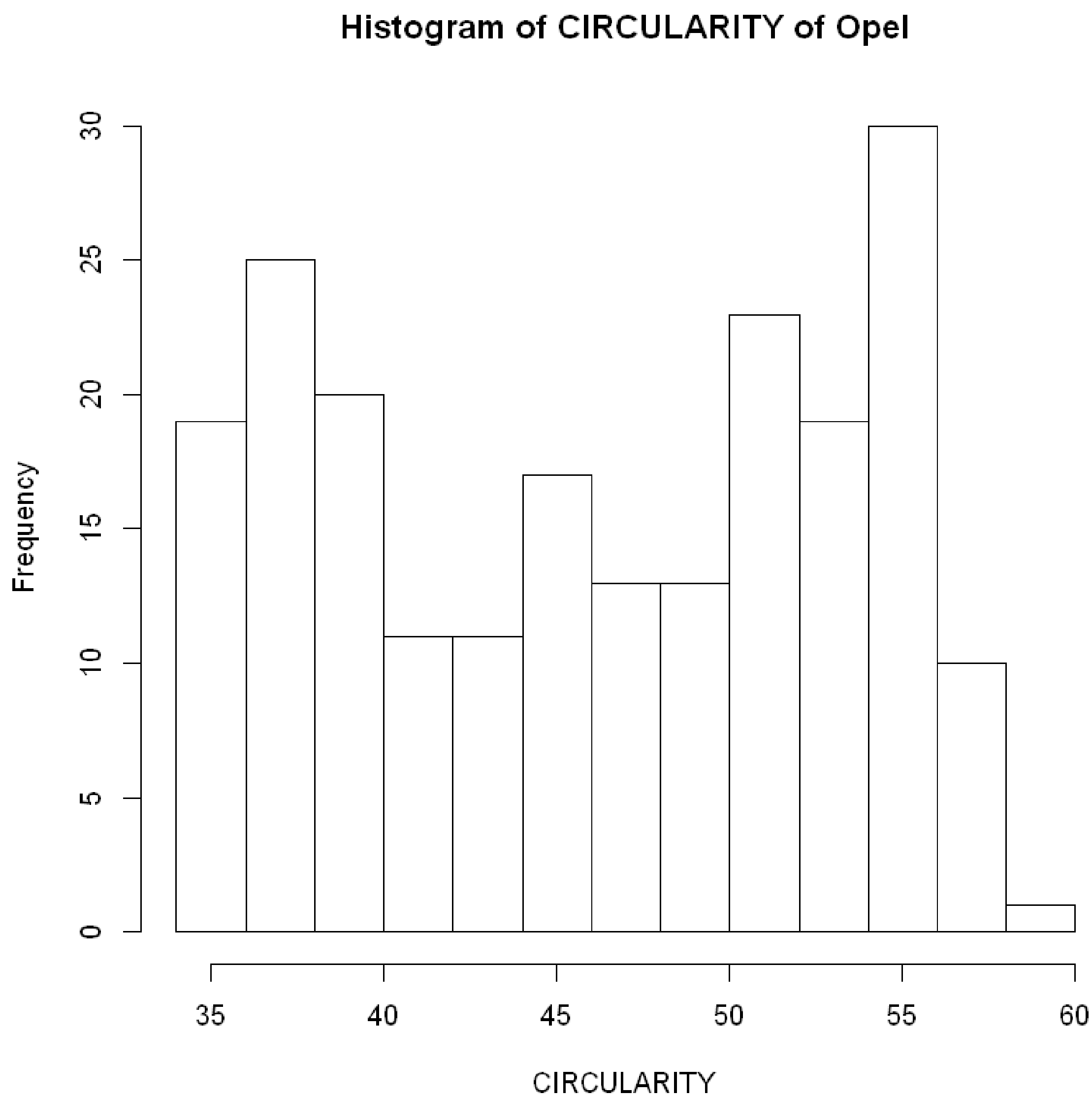
The histogram of Compactness for the whole dataset shows that the most frequency happens at the range from 85 to 90. The histogram of Compactness is skewed to the right and therefore there is more observation on the left side. The histogram of Circularity for the whole dataset shows the most occurrence happens at the range of 43 to 47. The histogram of Circularity is symmetrically skewed but also skewed to the left. This shows that the observation is more spread out and more frequency on the right side.

In [6]:
```
selected_data <- dataset[c(1,2,18,19)]
opel_hist_data <- selected_data[selected_data$V19 == "opel",]
hist(opel_hist_data$V1, main = "Histogram of COMPACTNESS of Opel", xlab = "COM
PACTNESS", ylab = "Frequency")
skewness(opel_hist_data$V1)
hist(opel_hist_data$V2, main = "Histogram of CIRCULARITY of Opel", xlab = "CIR
CULARITY", ylab = "Frequency")
skewness(opel_hist_data$V2)
```

-0.366577226551324



-0.112255583752042

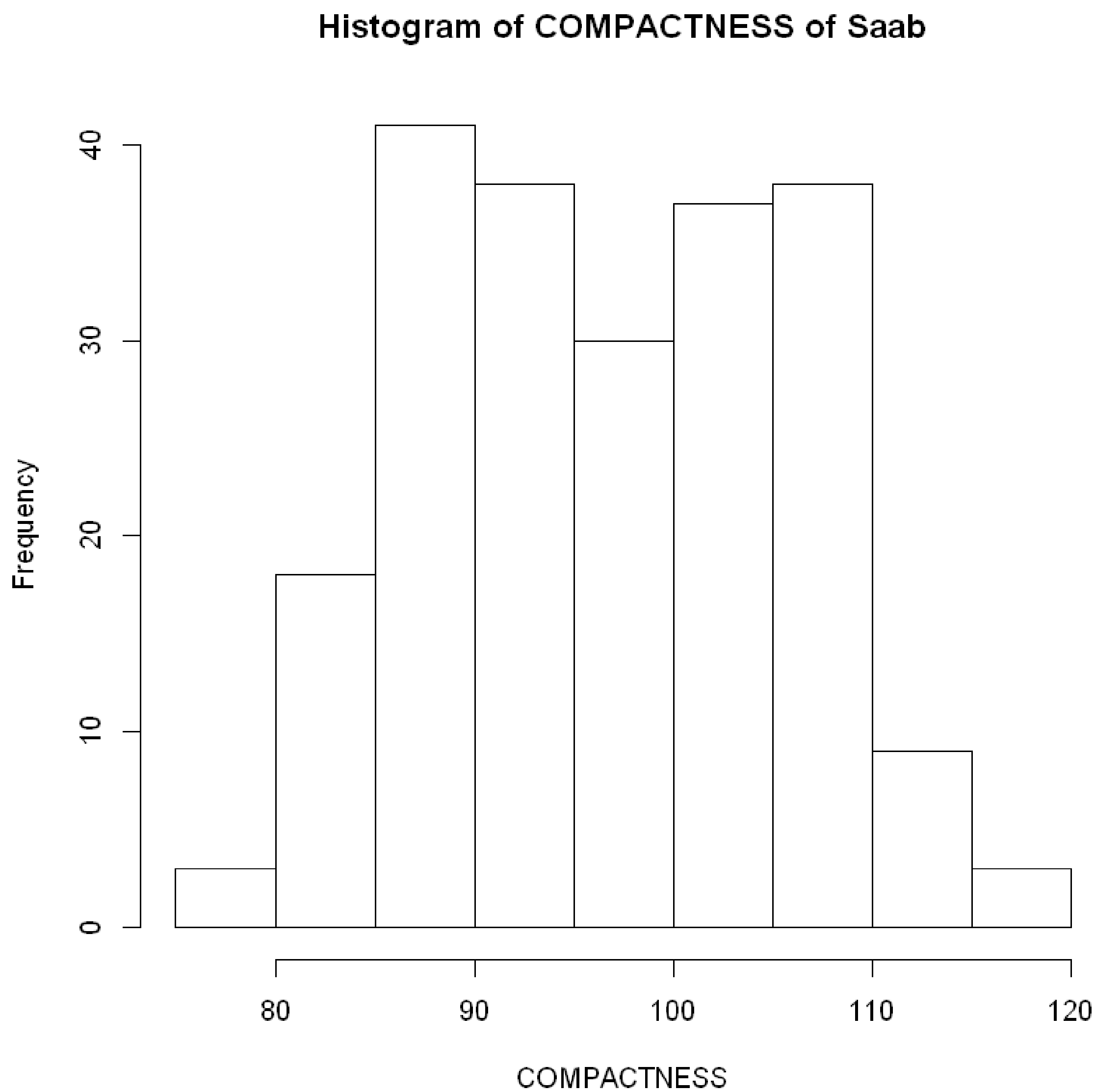


Analysis:

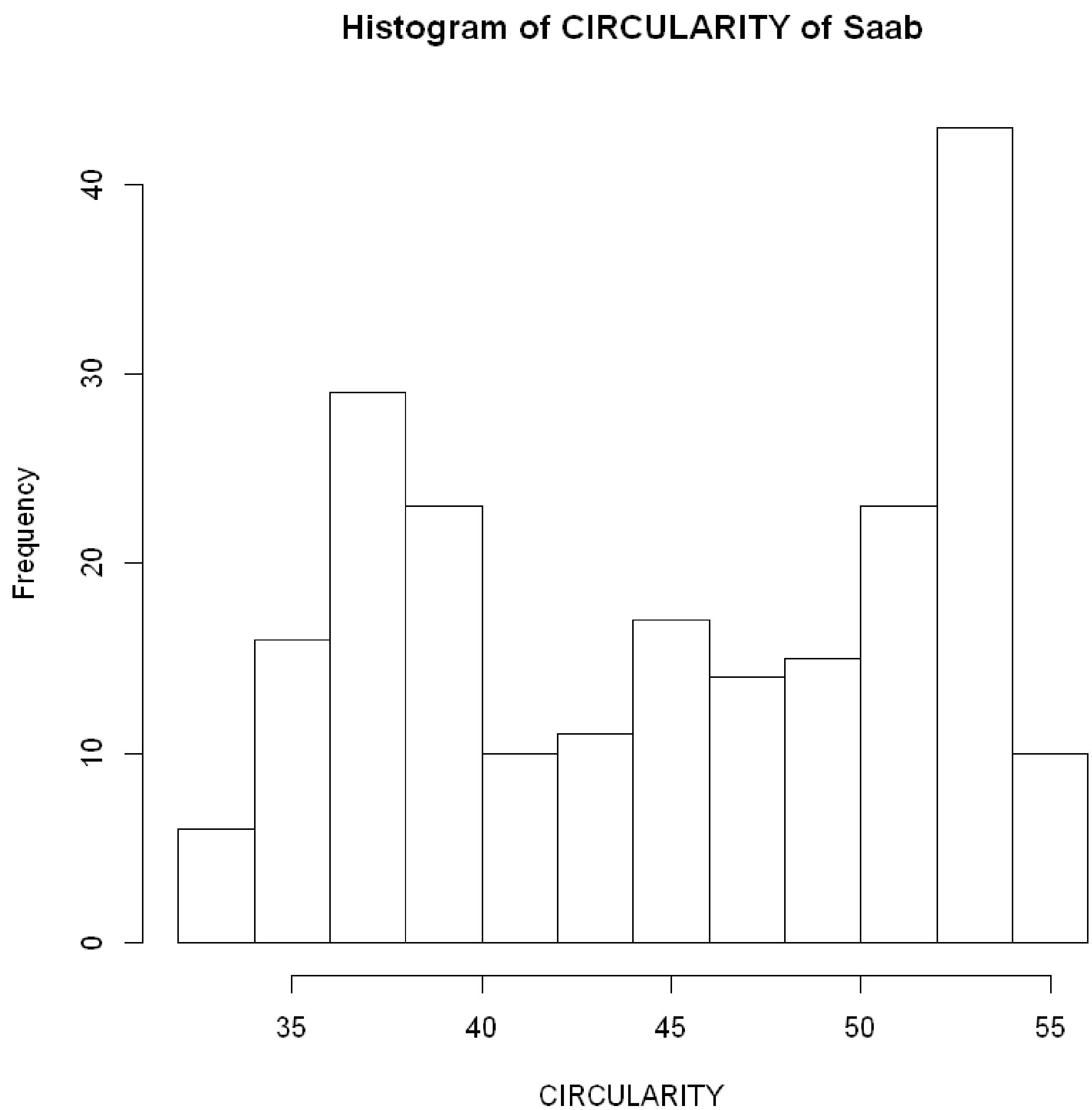
The histogram of Compactness for Opel shows that the most frequency happens at the range from 100 to 105. The histogram of Compactness for Opel is skewed to the left and therefore there is more observation on the right side. The histogram of Circularity for Opel shows its a bimodal and the two higest peak are between 36 to 38 and 53 to 56. The histogram of Circularity for Opel is bimodal skewed.

In [7]:
```
saab_hist_data <- selected_data[selected_data$V19 == 'saab',]
hist(saab_hist_data$V1, main = "Histogram of COMPACTNESS of Saab", xlab = "COM
PACTNESS", ylab = "Frequency")
skewness(saab_hist_data$V1)
hist(saab_hist_data$V2, , main = "Histogram of CIRCULARITY of Saab", xlab = "C
IRCULARITY", ylab = "Frequency")
skewness(saab_hist_data$V1)
```

0.0989796365630487



0.0989796365630487

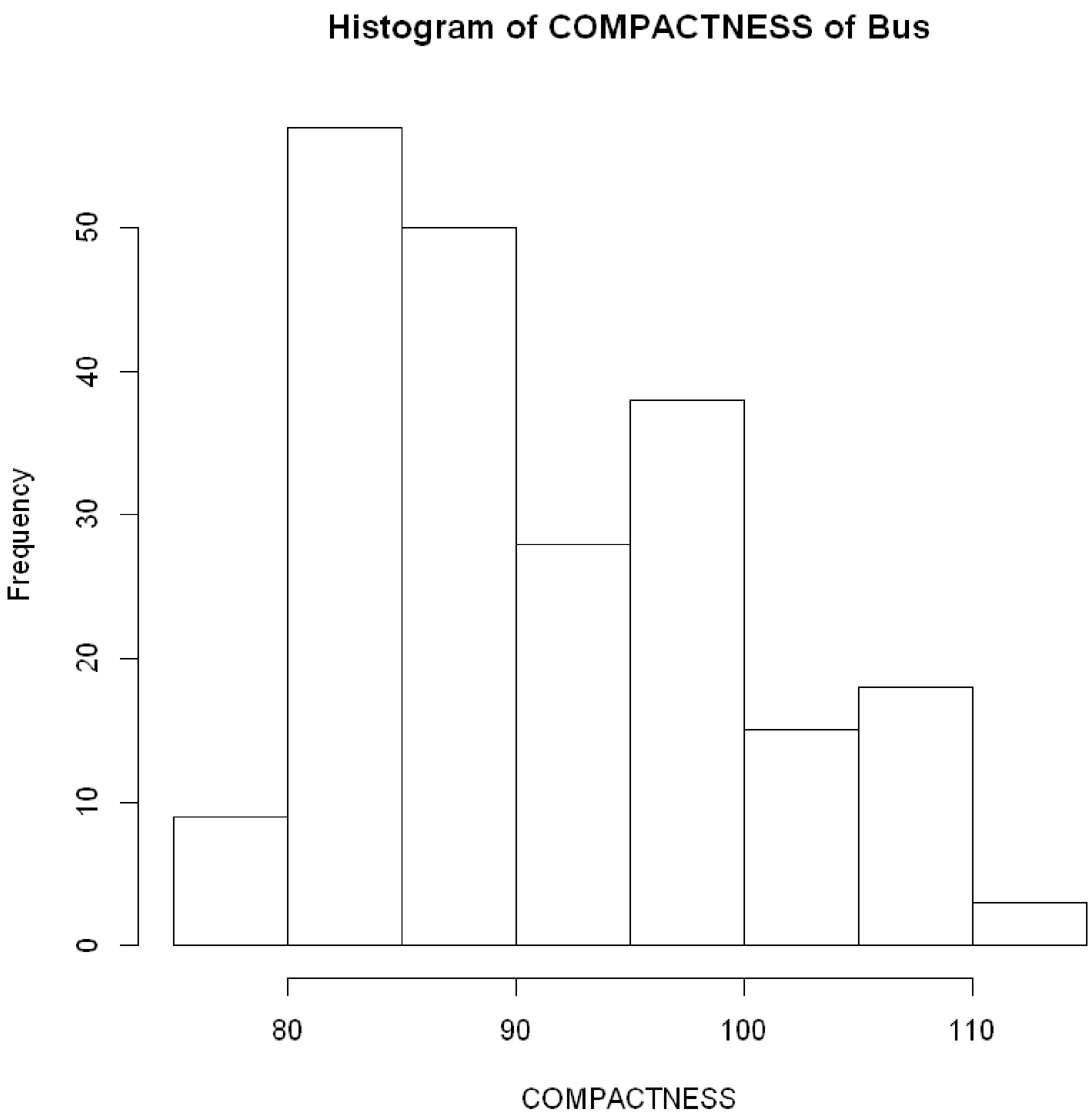


Analysis:

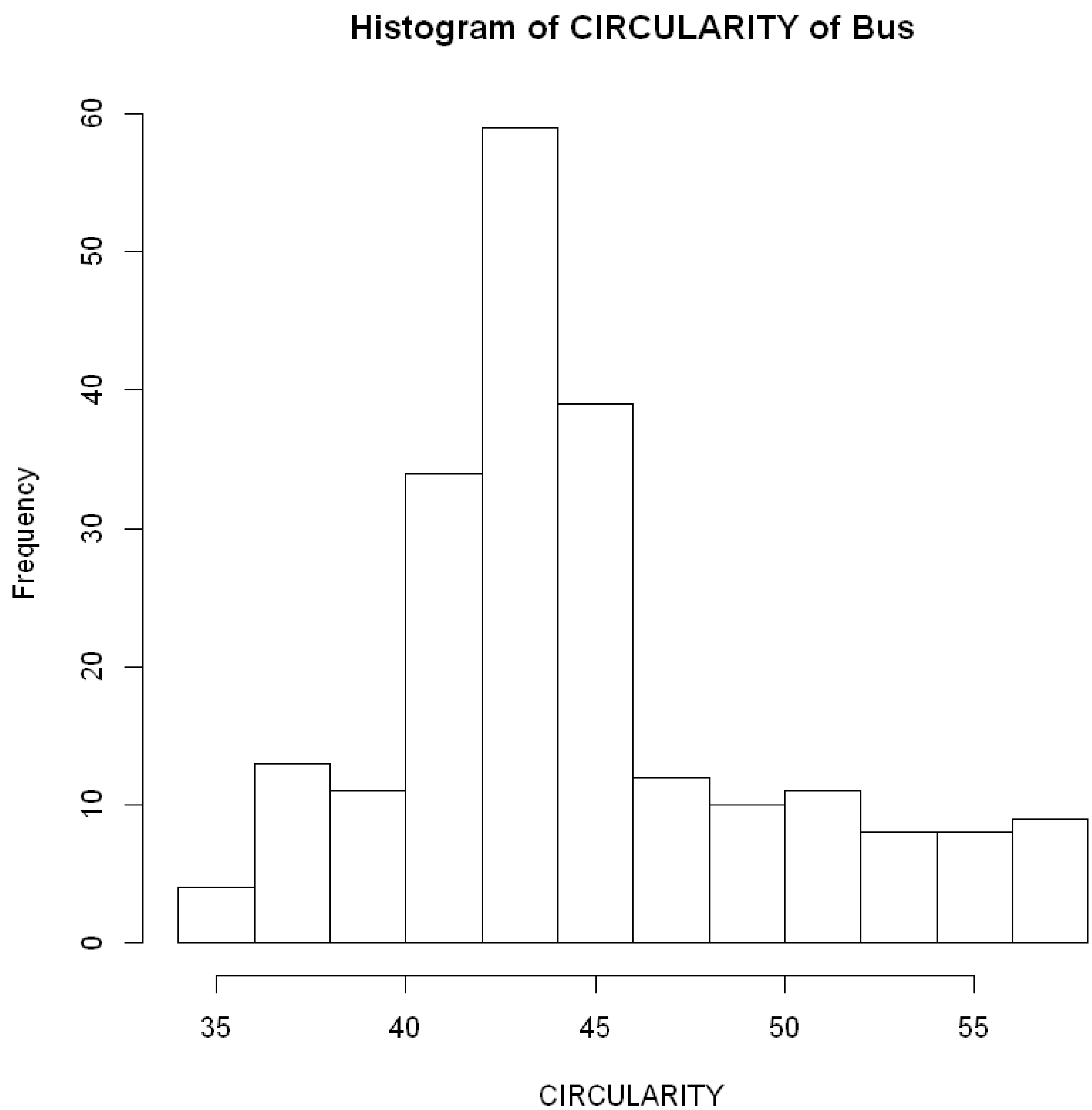
The histogram of Compactness for Saab shows that the most frequency happens at the range from 85 to 90 and 105 to 110. The histogram of Compactness for Saab is bimodal skewed. The histogram of Circularity for Saab shows its a bimodal and the two higest peak are between 36 to 38 and 53 to 54. The histogram of Circularity for Saab is bimodat skewed.

In [8]:
```
bus_hist_data <- selected_data[selected_data$V19 == 'bus',]
hist(bus_hist_data$V1, main = "Histogram of COMPACTNESS of Bus", xlab = "COMPA
CTNESS", ylab = "Frequency")
skewness(bus_hist_data$V1)
hist(bus_hist_data$V2, main = "Histogram of CIRCULARITY of Bus", xlab = "CIRCU
LARITY", ylab = "Frequency")
skewness(bus_hist_data$V2)
```

0.575407673996567



0.777615910060001

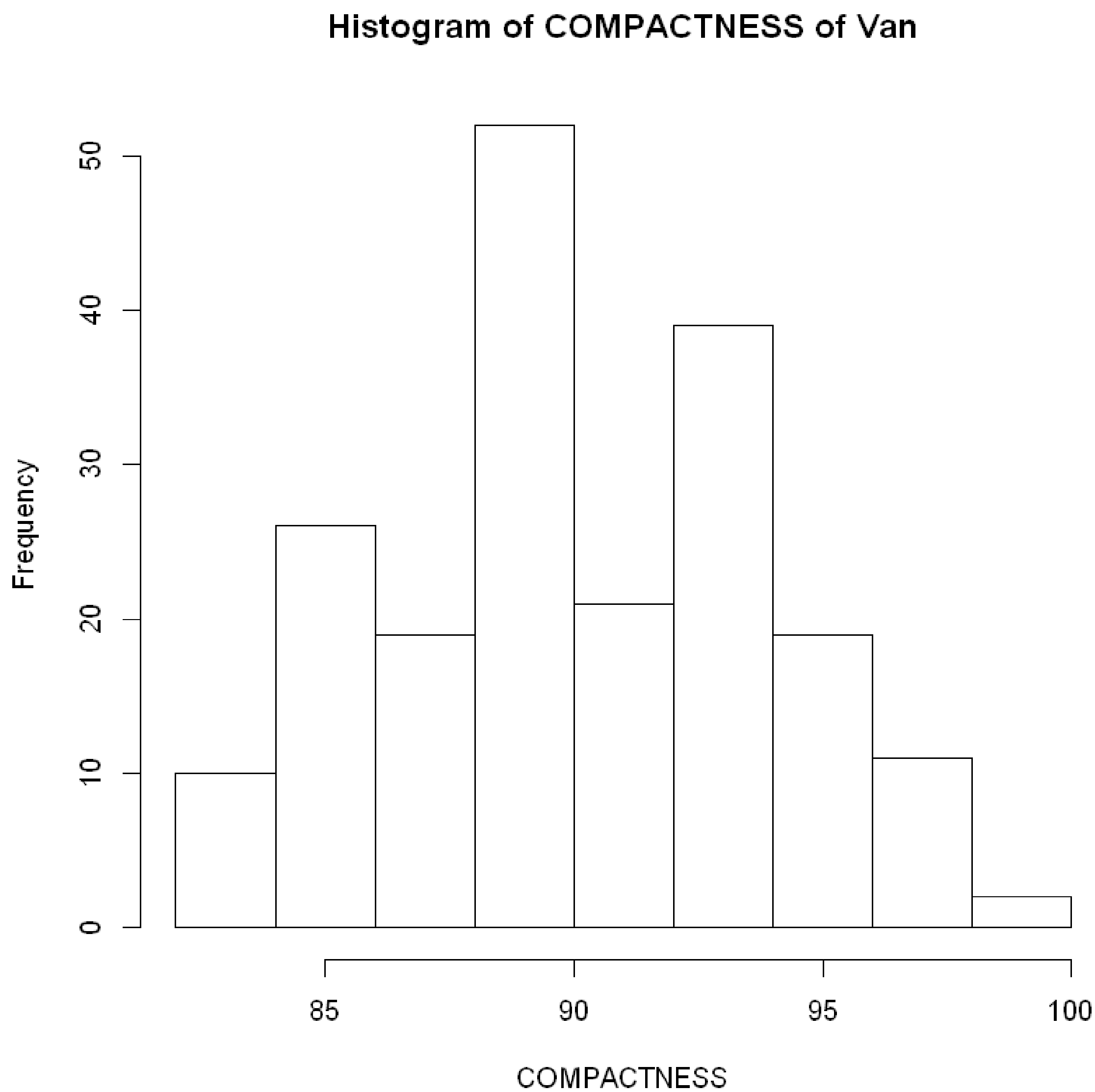


Analysis:

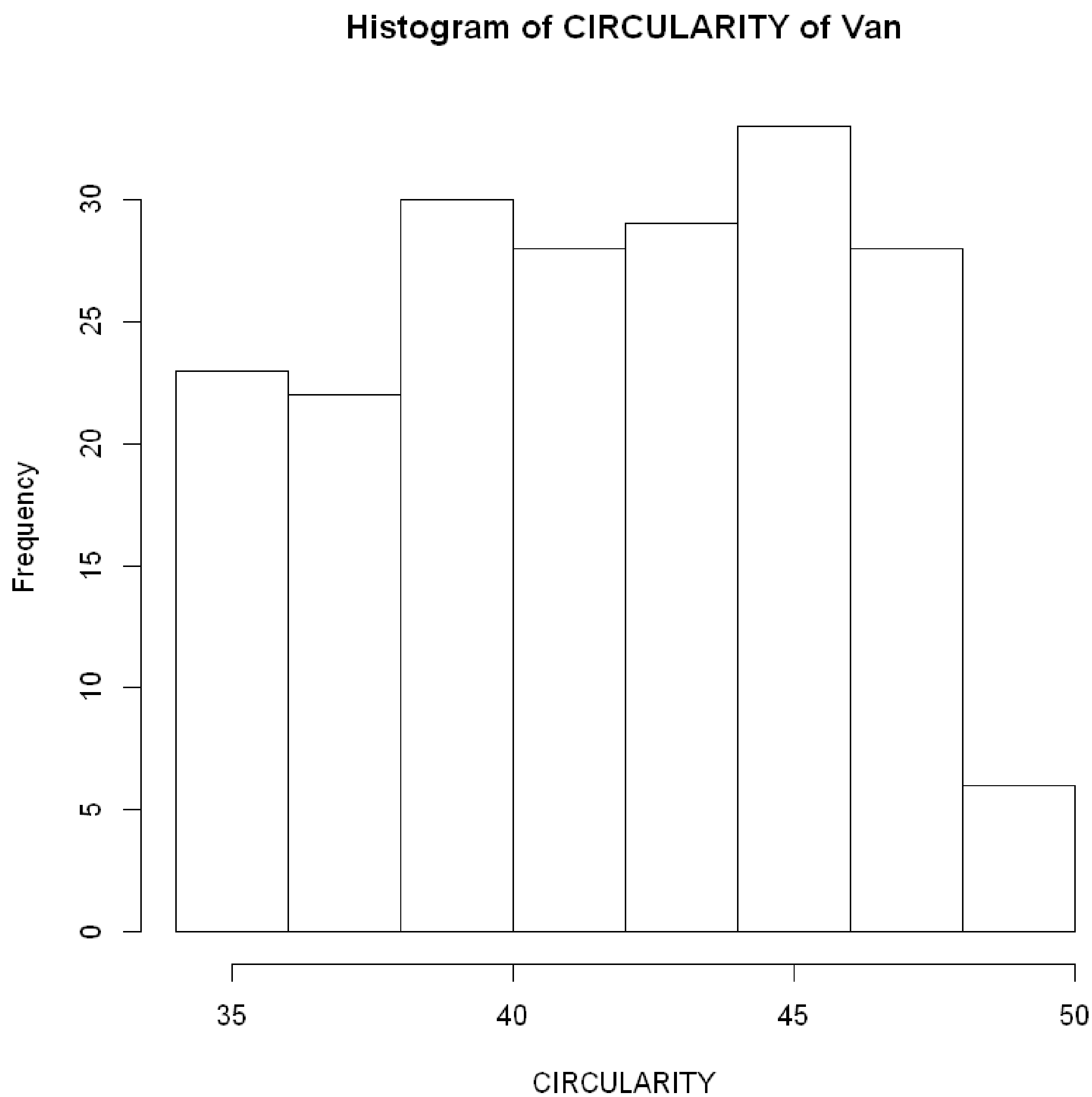
The histogram of Compactness for Bus shows that the most frequency happens at the range from 80 to 85 and 105 to 110. The histogram of Compactness for Bus is skewed to the right. The histogram of Circularity for Bus is skewed to the right

In [9]:
```
van_hist_data <- selected_data[selected_data$V19 == 'van',]
hist(van_hist_data$V1, main = "Histogram of COMPACTNESS of Van", xlab = "COMPA
CTNESS", ylab = "Frequency")
skewness(van_hist_data$V1)
hist(van_hist_data$V2, main = "Histogram of CIRCULARITY of Van", xlab = "CIRCU
LARITY", ylab = "Frequency")
skewness(van_hist_data$V2)
```

0.0770501309801016



-0.119511003689927

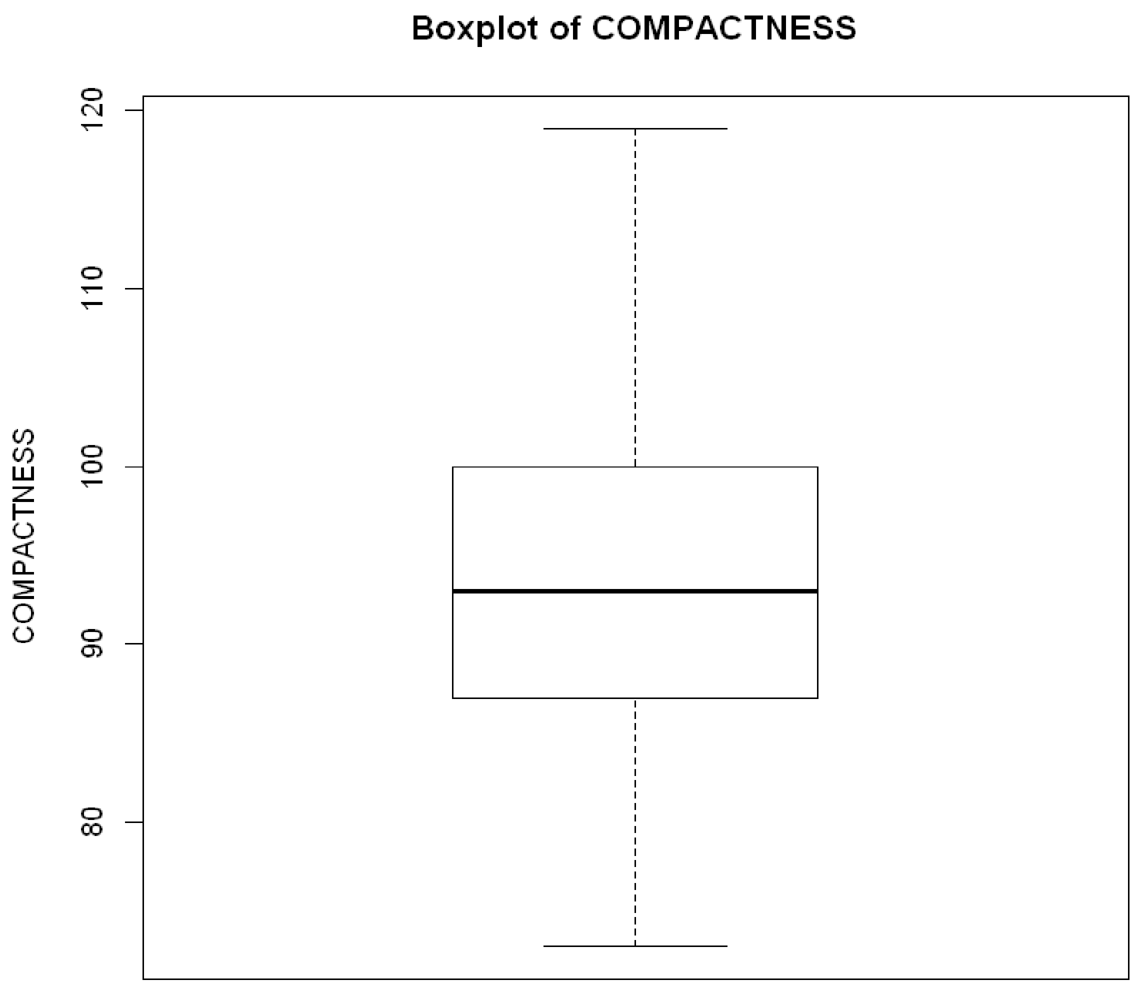


Analysis:

The histogram of Compactness for Van comb distribution. The histogram of Circularity for Bus has no pattern.

Number 4

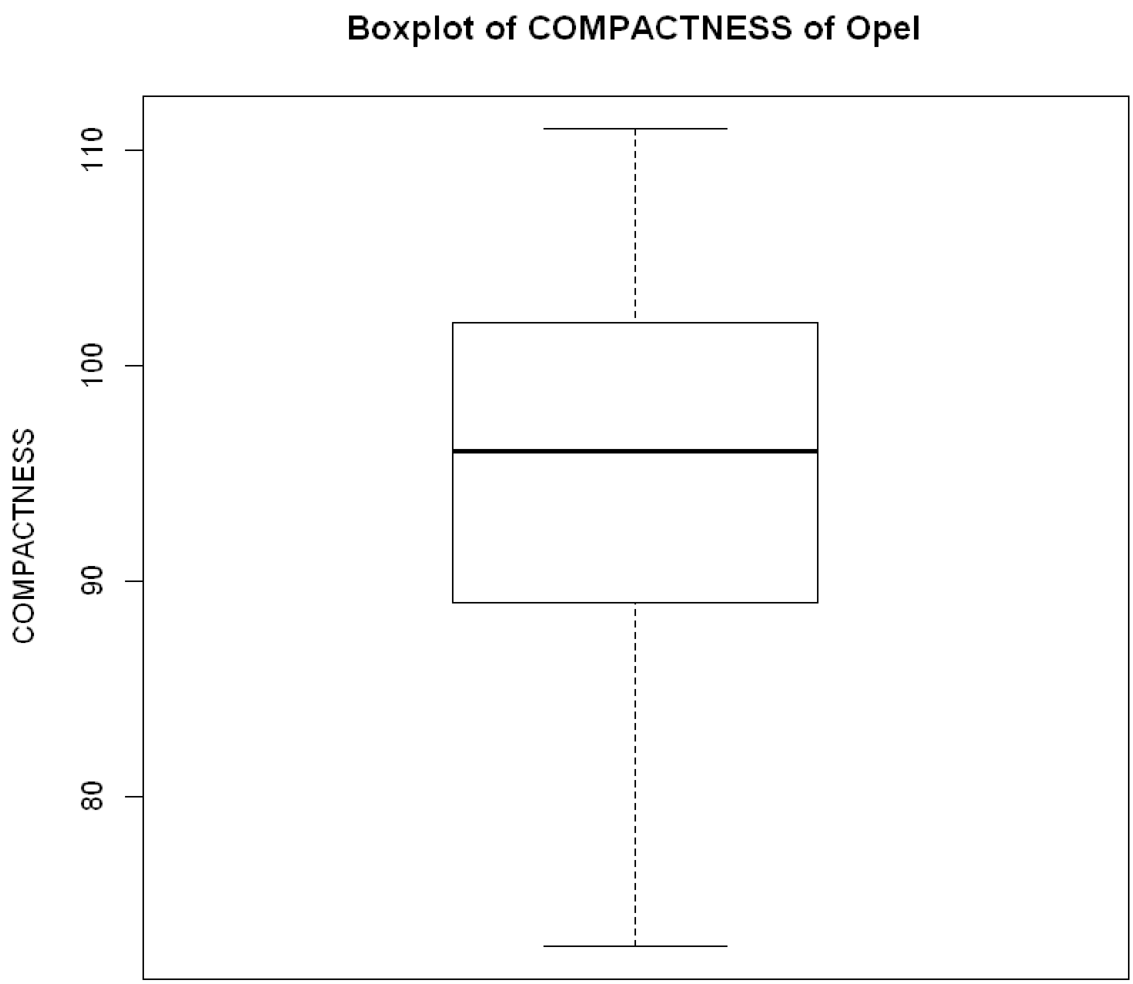
In [10]:
```
#Boxplot
boxplot(dataset$V1,main="Boxplot of COMPACTNESS", ylab = "COMPACTNESS")
```



Analysis:

Q1 and Q4 both vary in compactness but it seems that the compactness amongst all cars in Q4 all tend to have a longer range. In Q2 and Q3 they tend to vary a lot less with Q2 and Q3 hanging around the median. The cars tend to have a standard size.

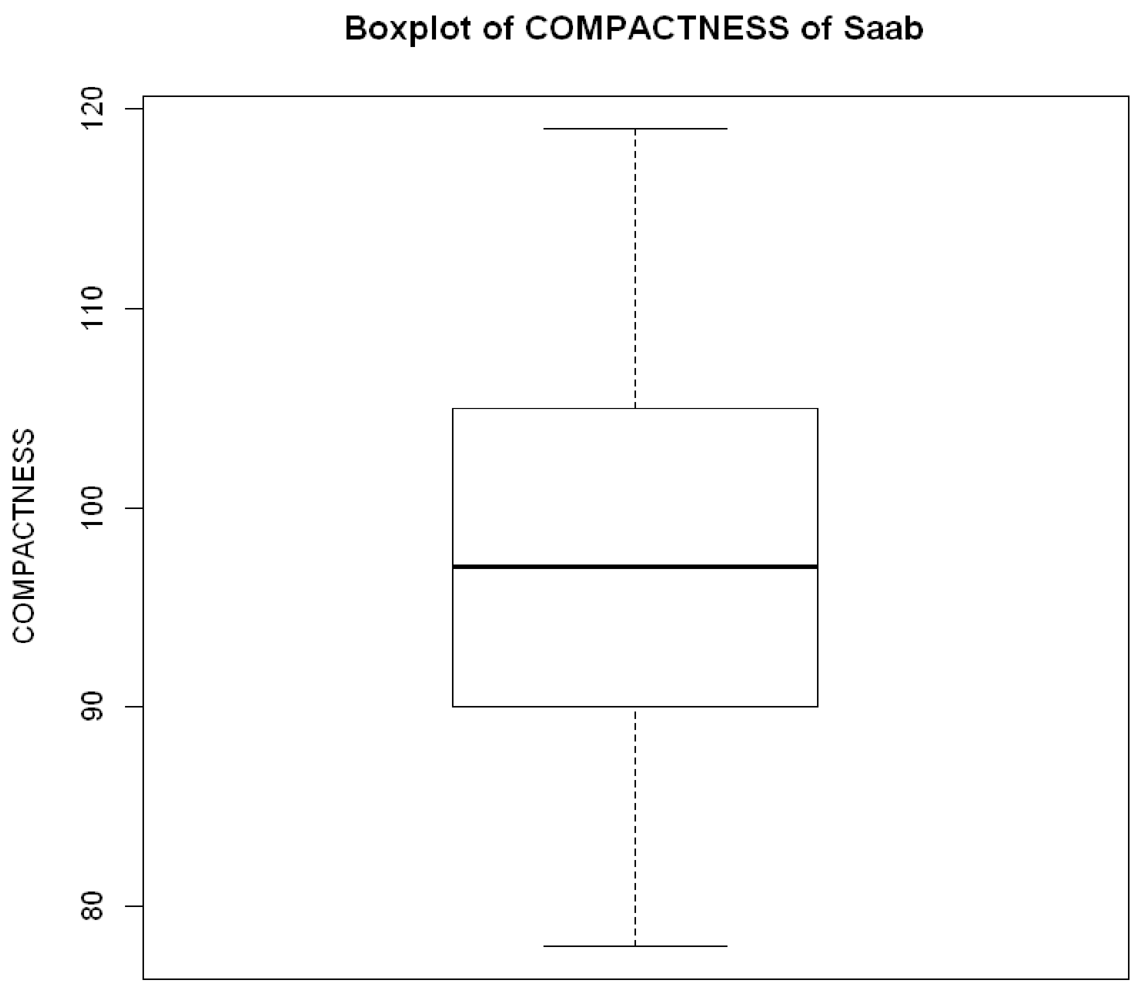
In [11]:
```
boxplot(opel_hist_data$V1,main="Boxplot of COMPACTNESS of Opel", ylab = "COMPA
CTNESS")
```



Analysis:

Quartile group 1 is extended. The compactness for opel has a wider range in the lower numbers. The positive quartile group also vary in compactness but not as much.

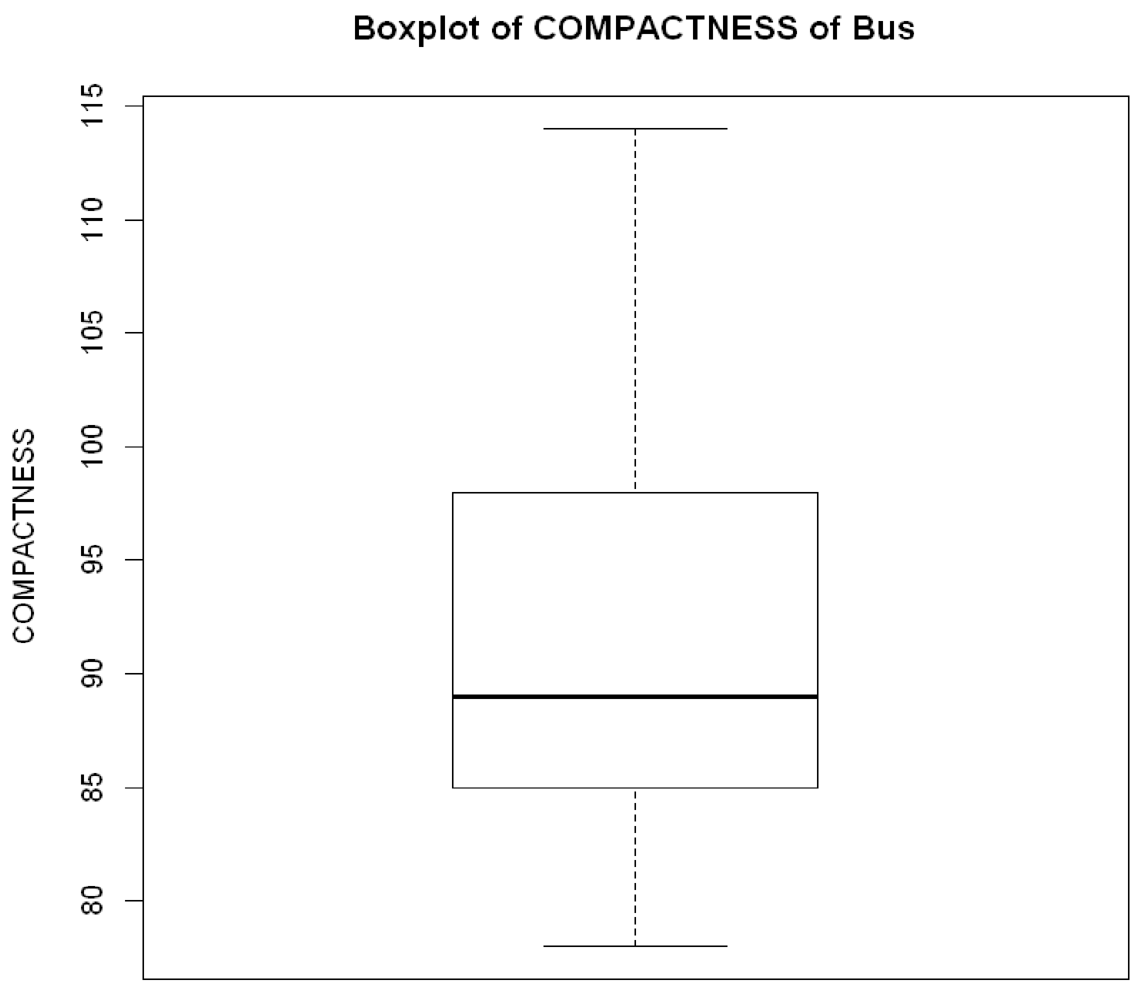
In [12]:
```
boxplot(saab_hist_data$V1,main="Boxplot of COMPACTNESS of Saab", ylab = "COMPACTNESS")
```



Analysis:

The compactness for Saab seems symmetrical with Q1 and Q4 ranging within the same size. Q2 and Q3 are split between the median about the same but with Q3 ranging slightly more

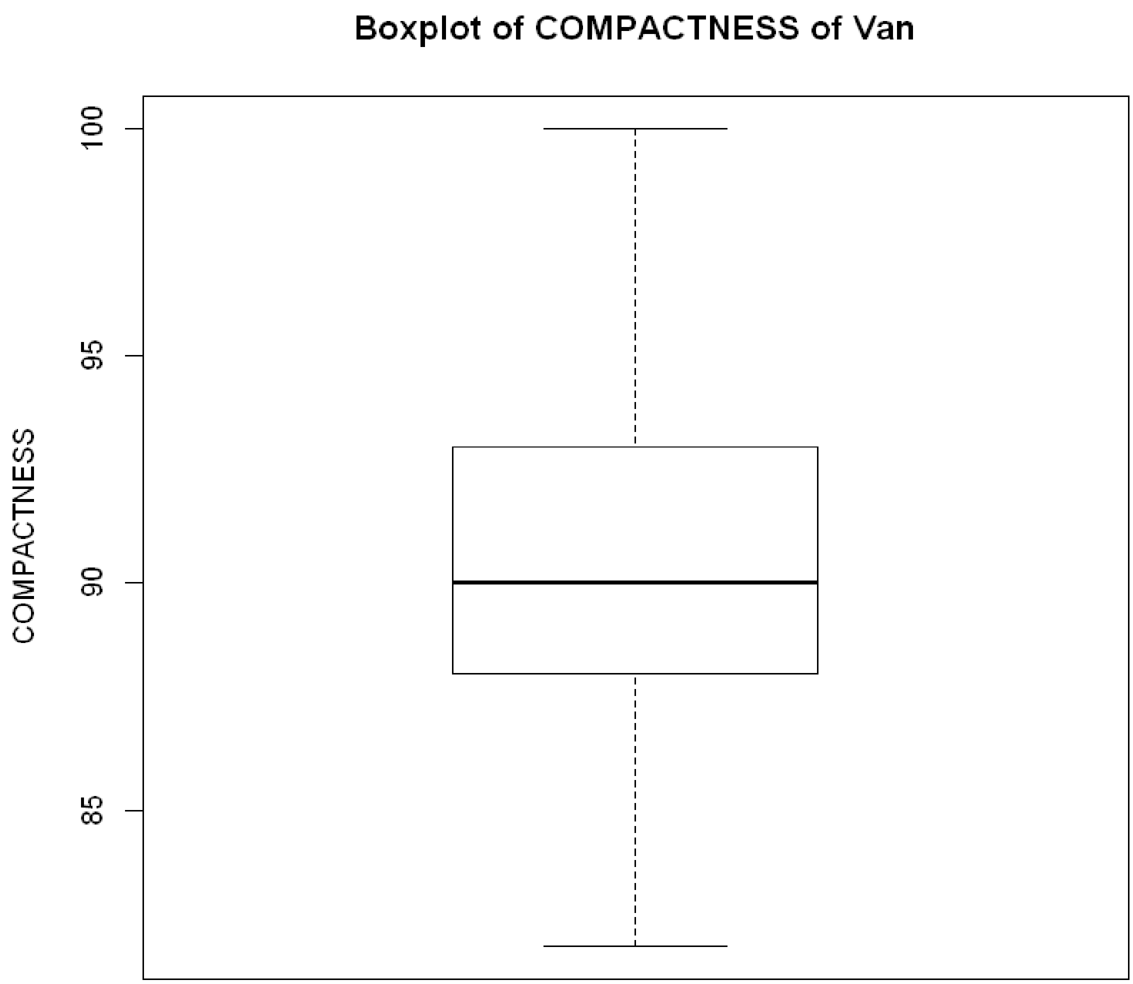
In [13]:
```
boxplot(bus_hist_data$V1,main="Boxplot of COMPACTNESS of Bus", ylab = "COMPACTNESS")
```



Analysis:

In Q1 and Q2 the compactness of bus seems to vary much less. Most buses have the same compactness around this range but past the median the compact of bus is vary way more in Q3 and Q4.

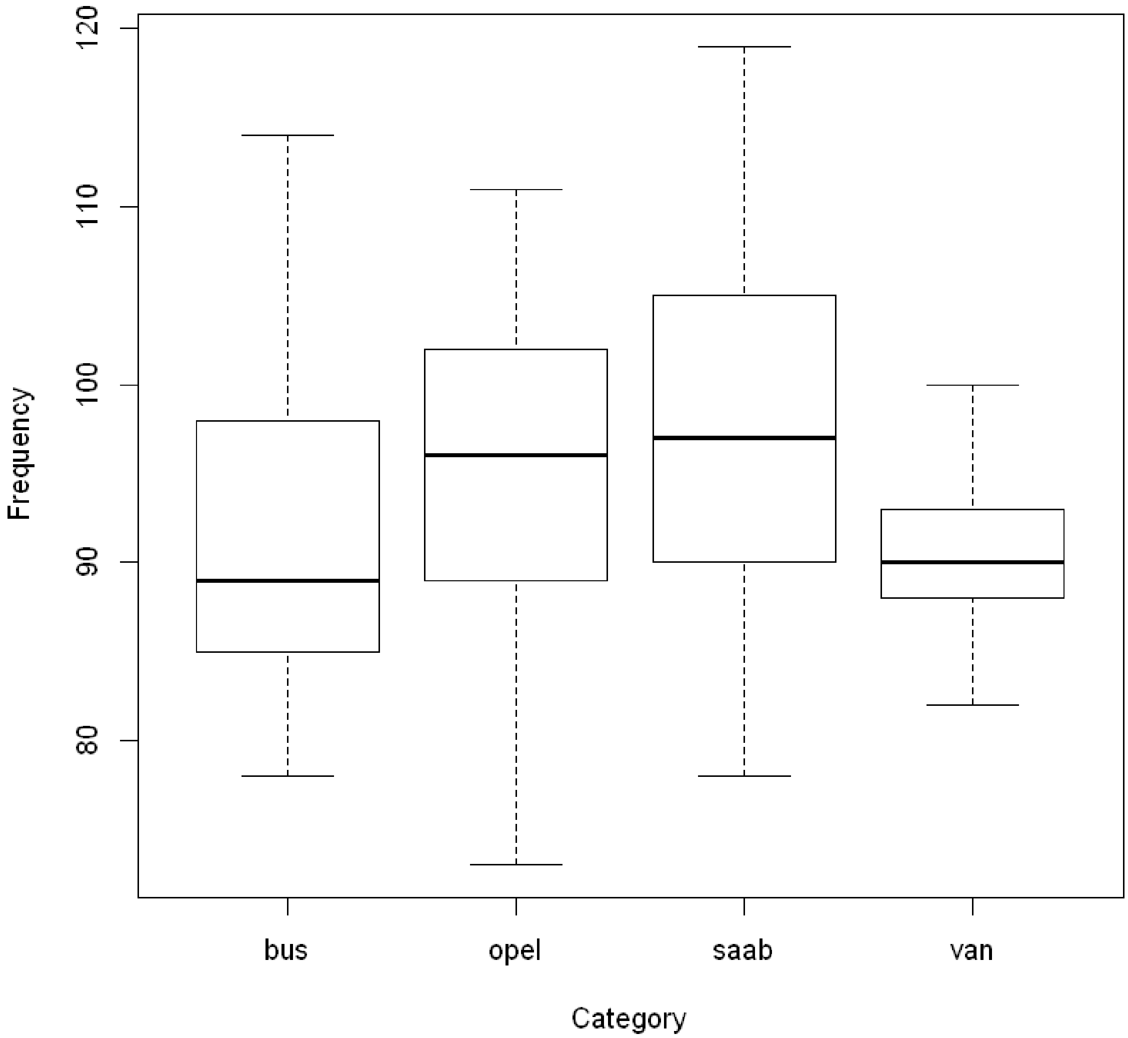
In [14]:
```
boxplot(van_hist_data$V1,main="Boxplot of COMPACTNESS of Van", ylab = "COMPACT
NESS")
```



Analysis

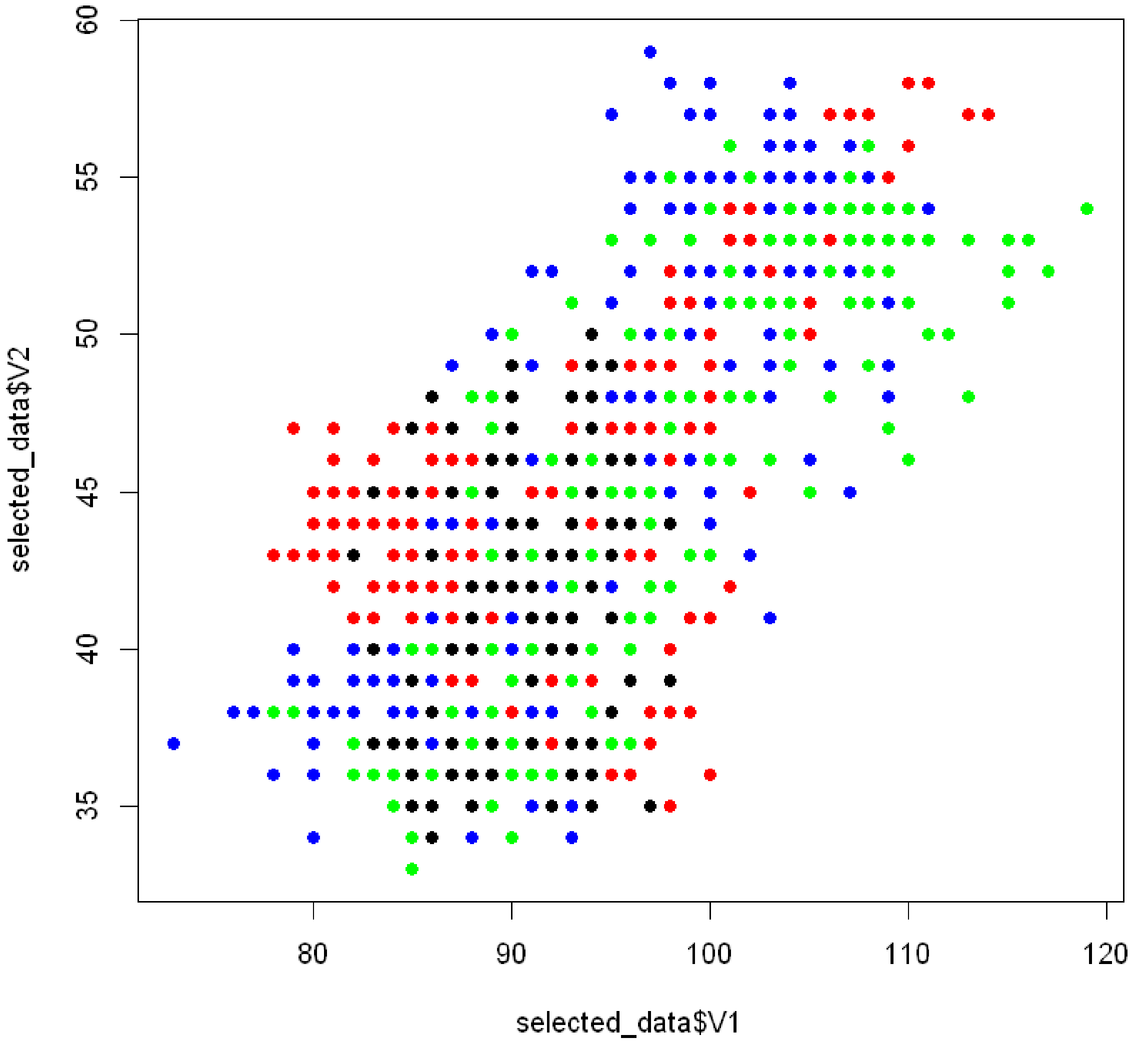
This boxplot seems very symmetrical. Q1 and Q4 are comparatively the same range and will vary a lot in compactness. In Q2 and Q3 the compact doesn’t as much as but does range more in Q3.

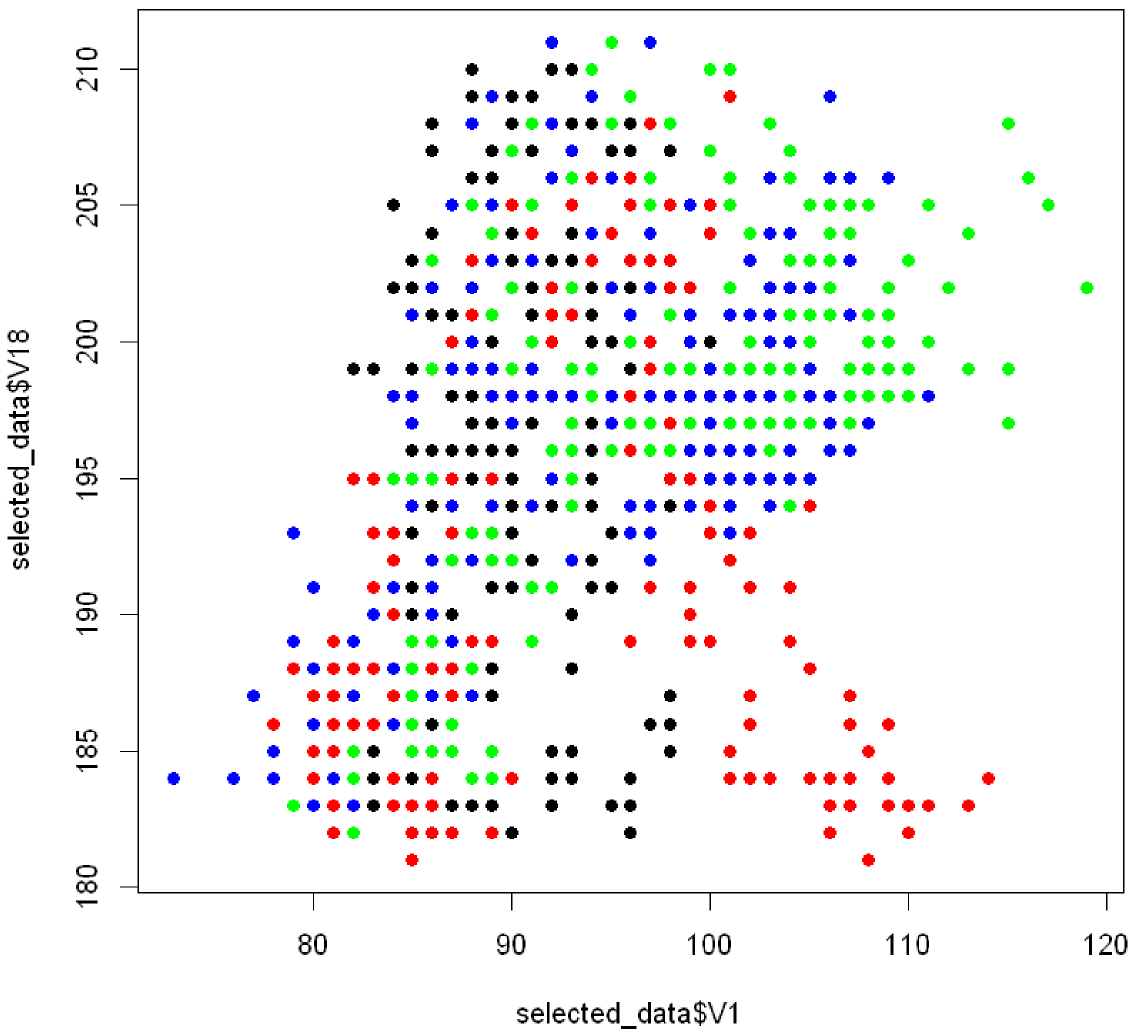
In [15]:
```
boxplot(V1~V19, data = selected_data, xlab = "Category", ylab = "Frequency")
```

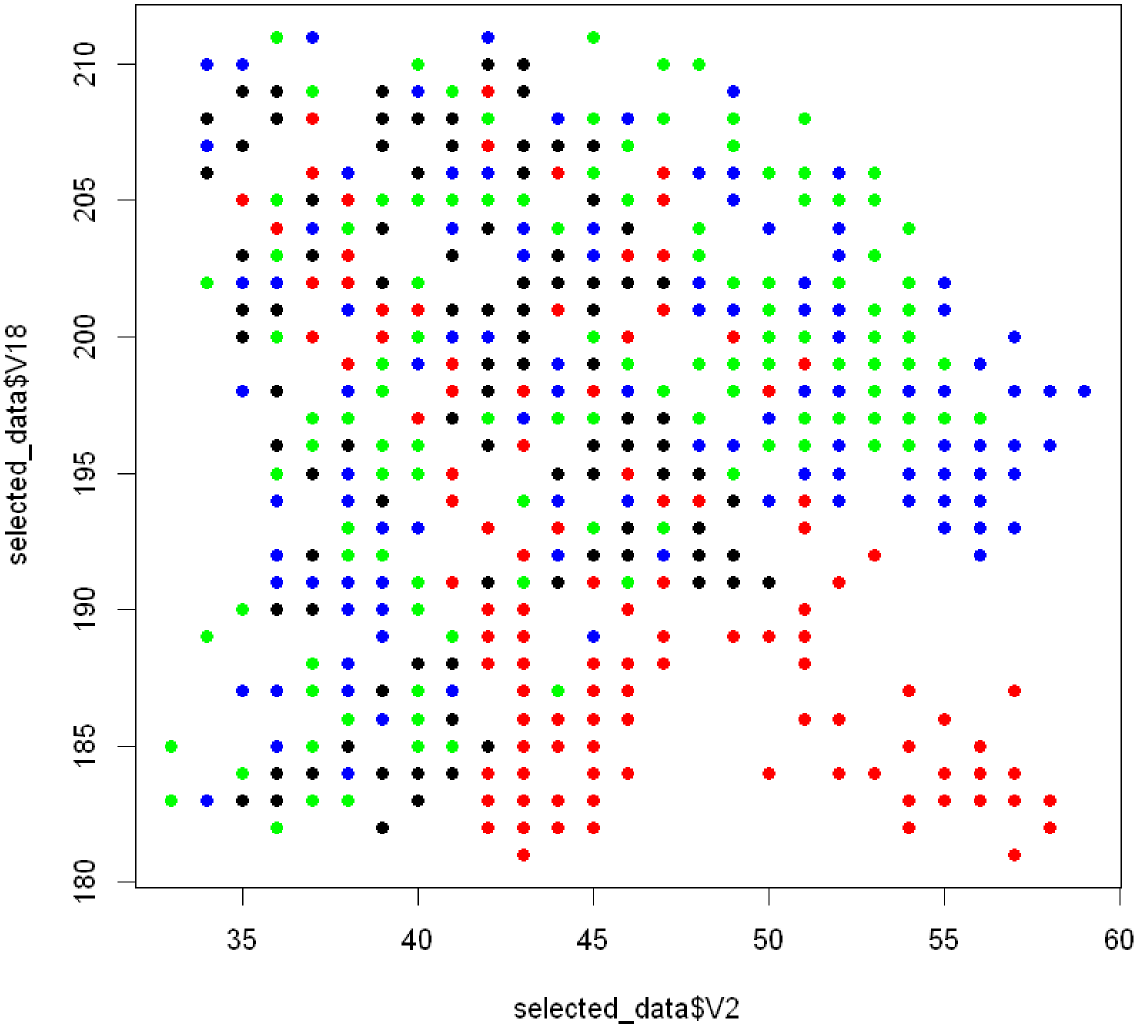


Number 5

In [16]:
```
# Supervised scatter
plot(selected_data$V1, selected_data$V2, col=c("red","blue","green","black")[dataset$V19],pch=19)
plot(selected_data$V1, selected_data$V18, col=c("red","blue","green","black")[dataset$V19],pch=19)
plot(selected_data$V2, selected_data$V18, col=c("red","blue","green","black")[dataset$V19],pch=19)
```


60
55
50
45
40
35
selected_data$V2
80
90
100
110
120
selected_data$V1


selected_data$V18
selected_data$V1
180
185
190
195
200
205
210
80
90
100
110
120



Analysis:

The scatter plot for Compactness VS. Circularity has a positive trend, by looking at the correlation between V1 and V2 is 0.6928692, which is a high correlation. The scatter plot for Compactness vs HOLLOWS RATIO has no correlation between V1 and V2. The correlation coefficient for V1 and V18 is 0.3655518, therefore it has a low correlation. The scatter plot for Circularity vs HOLLOWS RATIO has no correlation because the correlation coefficient is 0.03867702, which is very low.

Number 6

In [17]:
```
# 6a
# zscore
zscore <- dataset
zscore$V1 <- (dataset$V1 - mean(dataset$V1)) / sd(dataset$V1)
zscore$V2 <- (dataset$V2 - mean(dataset$V2)) / sd(dataset$V2)
zscore$V3 <- (dataset$V3 - mean(dataset$V3)) / sd(dataset$V3)
zscore$V4 <- (dataset$V4 - mean(dataset$V4)) / sd(dataset$V4)
zscore$V5 <- (dataset$V5 - mean(dataset$V5)) / sd(dataset$V5)
zscore$V6 <- (dataset$V6 - mean(dataset$V6)) / sd(dataset$V6)
zscore$V7 <- (dataset$V7 - mean(dataset$V7)) / sd(dataset$V7)
zscore$V8 <- (dataset$V8 - mean(dataset$V8)) / sd(dataset$V8)
zscore$V9 <- (dataset$V9 - mean(dataset$V9)) / sd(dataset$V9)

zscore$V10 <- (dataset$V10 - mean(dataset$V10)) / sd(dataset$V10)
zscore$V11 <- (dataset$V11 - mean(dataset$V11)) / sd(dataset$V11)
zscore$V12 <- (dataset$V12 - mean(dataset$V12)) / sd(dataset$V12)
zscore$V13 <- (dataset$V13 - mean(dataset$V13)) / sd(dataset$V13)
zscore$V14 <- (dataset$V14 - mean(dataset$V14)) / sd(dataset$V14)
zscore$V15 <- (dataset$V15 - mean(dataset$V15)) / sd(dataset$V15)
zscore$V16 <- (dataset$V16 - mean(dataset$V16)) / sd(dataset$V16)
zscore$V17 <- (dataset$V17 - mean(dataset$V17)) / sd(dataset$V17)
zscore$V18 <- (dataset$V18 - mean(dataset$V18)) / sd(dataset$V18)
```

In [18]:
```
# 6b
zscore$B[zscore$V19== 'bus'] <- 1
zscore$B[zscore$V19== 'opel'] <- 0
zscore$B[zscore$V19== 'saab'] <- 0
zscore$B[zscore$V19== 'van'] <- 0

# 6c
zscore$V[zscore$V19== 'bus'] <- 0
zscore$V[zscore$V19== 'opel'] <- 0
zscore$V[zscore$V19== 'saab'] <- 0
zscore$V[zscore$V19== 'van'] <- 1
```

In [19]:
```
modelB <- lm(B~(V1+V2+V3+V4+V5+V6+V7+V8+V9+V10+V11+V12+V13+V14+V15+V16+V17+V18
), data=zscore)
```

In [20]:
```
modelV <- lm(V~(V1+V2+V3+V4+V5+V6+V7+V8+V9+V10+V11+V12+V13+V14+V15+V16+V17+V18
), data=zscore)
```

In [21]:
```
# summary about the linear model, includes R square
summary(modelB)
summary(modelV)
```

```
Call:
lm(formula = B ~ (V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 +
    V10 + V11 + V12 + V13 + V14 + V15 + V16 + V17 + V18), data = zscore)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.73691 -0.16726 -0.01646  0.15295  0.83939 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.257683   0.008356  30.840  < 2e-16 ***
V1          -0.018945   0.019507  -0.971  0.33172    
V2           0.191161   0.060316   3.169  0.00158 ** 
V3          -0.167171   0.031621  -5.287 1.59e-07 ***
V4          -0.776887   0.052329 -14.846  < 2e-16 ***
V5           0.527168   0.034175  15.425  < 2e-16 ***
V6          -0.111289   0.017127  -6.498 1.41e-10 ***
V7          -0.115981   0.351989  -0.330  0.74186    
V8          -0.634434   0.104056  -6.097 1.66e-09 ***
V9          -0.130120   0.080091  -1.625  0.10462    
V10         -0.124176   0.043533  -2.852  0.00445 ** 
V11         -0.024460   0.057135  -0.428  0.66869    
V12          0.321222   0.253045   1.269  0.20465    
V13          0.013634   0.030918   0.441  0.65936    
V14          0.034831   0.029339   1.187  0.23550    
V15         -0.047227   0.009686  -4.876 1.30e-06 ***
V16          0.052184   0.010846   4.811 1.78e-06 ***
V17          0.409145   0.034028  12.024  < 2e-16 ***
V18         -0.329291   0.034278  -9.606  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.243 on 827 degrees of freedom
Multiple R-squared:  0.6982,    Adjusted R-squared:  0.6916 
F-statistic: 106.3 on 18 and 827 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = V ~ (V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 +
    V10 + V11 + V12 + V13 + V14 + V15 + V16 + V17 + V18), data = zscore)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67312 -0.16842 -0.00961  0.16415  0.88711

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.235225   0.008485  27.723  < 2e-16 ***
V1           0.155084   0.019808   7.829 1.50e-14 ***
V2          -0.568889   0.061248  -9.288  < 2e-16 ***
V3           0.378453   0.032110  11.786  < 2e-16 ***
V4          -0.274483   0.053138  -5.166 3.01e-07 ***
V5           0.152634   0.034703   4.398 1.23e-05 ***
V6          -0.039018   0.017392  -2.243 0.025135 *
V7          -1.362382   0.357428  -3.812 0.000148 ***
V8           0.230107   0.105664   2.178 0.029708 *
V9          -0.091591   0.081329  -1.126 0.260413
V10          0.677635   0.044206  15.329  < 2e-16 ***
V11          0.072552   0.058018   1.251 0.211463
V12          1.056310   0.256955   4.111 4.33e-05 ***
V13          0.037094   0.031396   1.181 0.237751
V14          0.078755   0.029793   2.643 0.008362 **
V15         -0.034964   0.009835  -3.555 0.000400 ***
V16         -0.040815   0.011014  -3.706 0.000225 ***
V17         -0.001221   0.034554  -0.035 0.971826
V18          0.081658   0.034808   2.346 0.019212 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2468 on 827 degrees of freedom
Multiple R-squared:  0.669,     Adjusted R-squared:  0.6618
F-statistic: 92.88 on 18 and 827 DF,  p-value: < 2.2e-16
```

Analysis:

- The z score coefficients tell us how closely each attribute can be tied back to the instance being a bus or a van depending on the model that you use. The negative means that whatever attribute has a negative z score doesn't reflect the attribute of the model. A positive coefficient means that whatever attribute has a positive z score has a higher of it being the model's attribute. The closer the coefficient is to +1 means that the closer the attribute is to it being a model attribute.
- For bus, Attribute V17 has the highest chance of it being a bus attribute with a ~41% rate. Meanwhile, Attribute V8 has the worst chance of it being a bus attribute with a negative ~63% rate.
- For van, Attribute V10 has the highest chance of it being a van attribute with a ~68% rate. Meanwhile, Attribute V7 has the worst chance of it being a van attribute with a ~136% rate.
- RSquare represents how well the entire model is fitting as bus attributes. In our case, roughly 70% of the variance of the bus model found in the response variable can be tied back to it being a bus attribute. Furthermore, roughly 67% of the variance of the van model found in the response variable can be tied back to it being a van attribute.
- Regarding to what extent the coefficient agrees with each other, we see a similarity in the intercept V7, V9, and V15.

Number 7

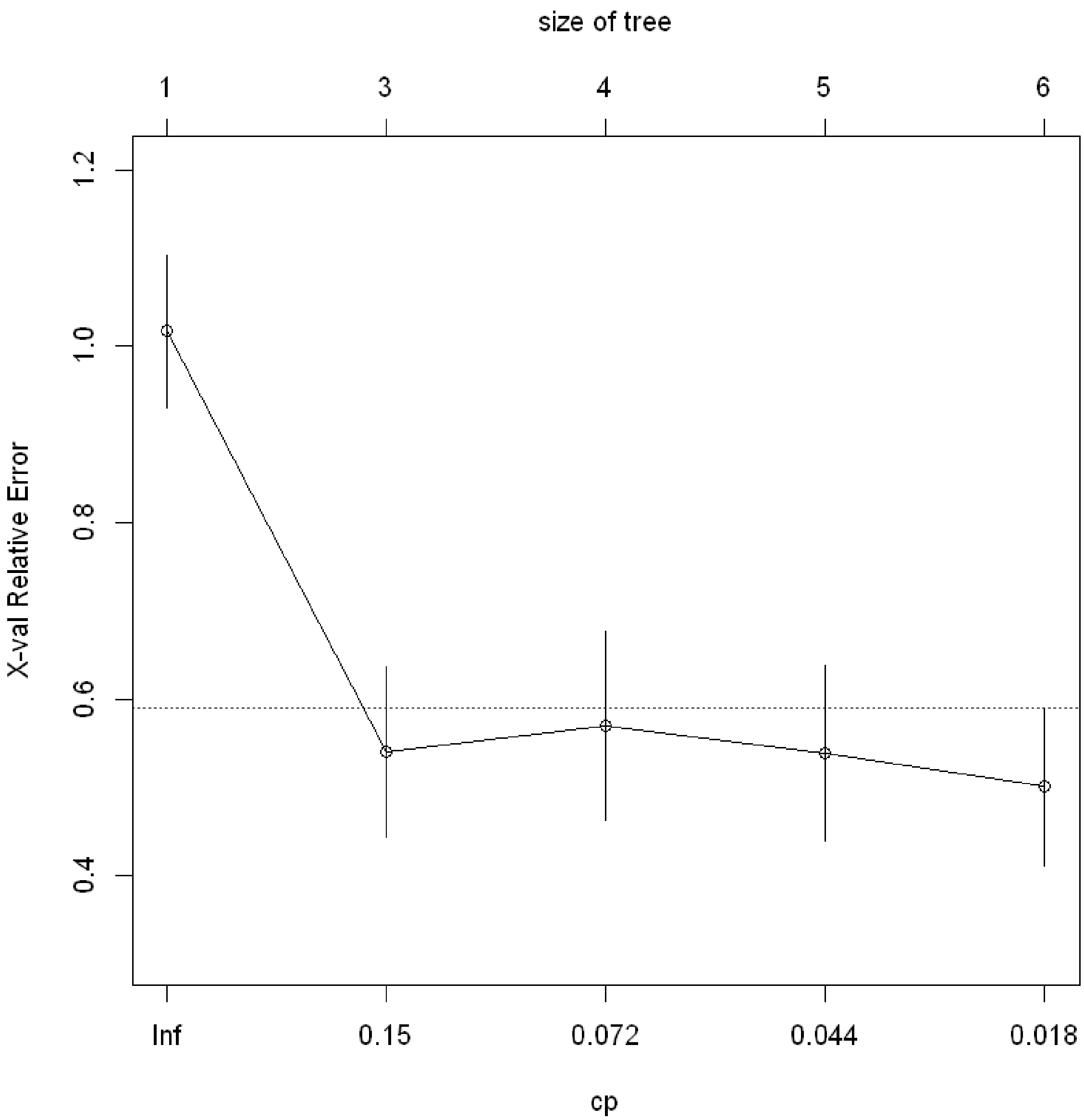
In [22]:
```
bus_data <- select(zscore,V1,V2,V3,V4,V5,V6,V7,V8,V9,V10,V11,V12,V13,V14,V15,V
16,V17,V18, B)
set.seed(123)
sample <- sample.split(bus_data$B, SplitRatio=0.8)

bus_train = subset(bus_data, sample == TRUE)
bus_test = subset(bus_data, sample == FALSE)
```

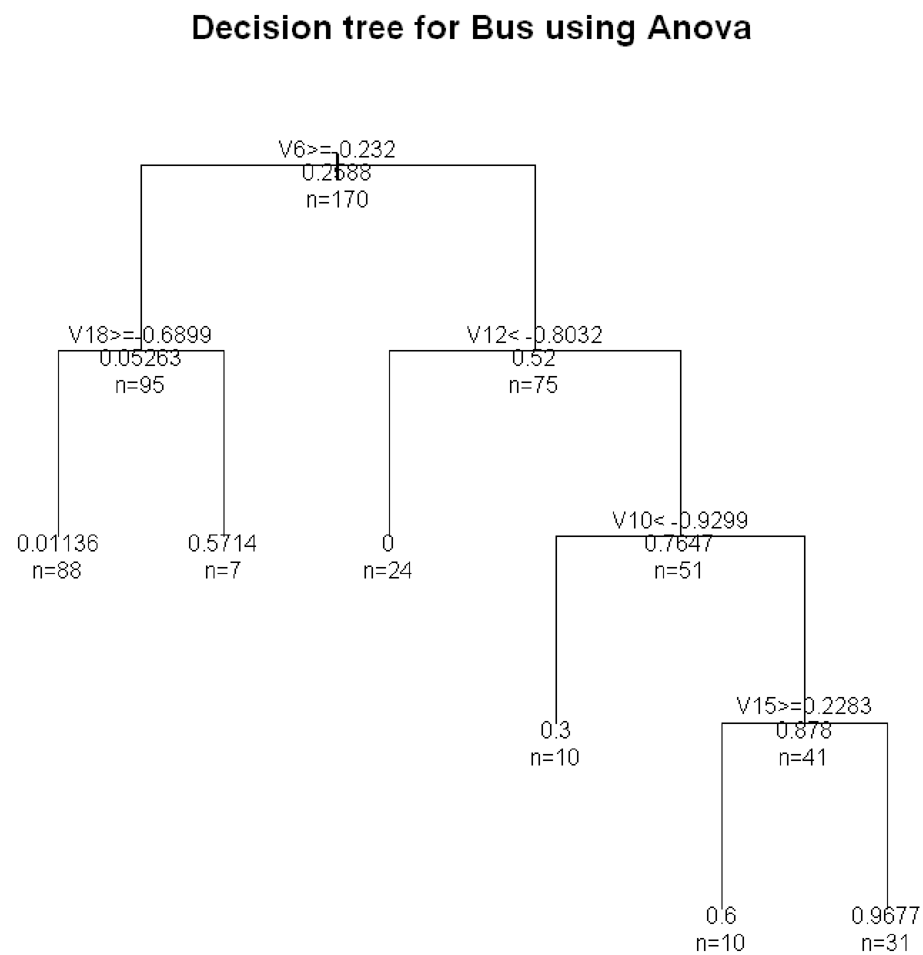
In [23]:
```
fit <-rpart(B~.,
            method="anova", data=bus_test)
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits
plot(fit, uniform=TRUE,
     main="Decision tree for Bus using Anova",margin=0.05)
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

```
Regression tree:
rpart(formula = B ~ ., data = bus_test, method = "anova")

Variables actually used in tree construction:
[1] V10 V12 V15 V18 V6

Root node error: 32.612/170 = 0.19183

n= 170

        CP nsplit rel error  xerror     xstd
1 0.286683      0   1.00000 1.01709 0.085927
2 0.082370      2   0.42663 0.54112 0.096189
3 0.062368      3   0.34426 0.56993 0.107004
4 0.031354      4   0.28190 0.53857 0.099394
5 0.010000      5   0.25054 0.50106 0.089167
Call:
rpart(formula = B ~ ., data = bus_test, method = "anova")
  n= 170

          CP nsplit rel error    xerror       xstd
1 0.28668262      0 1.0000000 1.0170933 0.08592686
2 0.08236987      2 0.4266348 0.5411228 0.09618903
3 0.06236768      3 0.3442649 0.5699257 0.10700380
4 0.03135378      4 0.2818972 0.5385658 0.09939388
5 0.01000000      5 0.2505434 0.5010609 0.08916687

Variable importance
V12  V3  V6 V11  V7  V8  V9 V10 V18  V1  V2 V13 V14  V4 V16 V15  V5
 14  10  10  10  10  10   8   7   7   4   2   2   1   1   1   1   1

Node number 1: 170 observations,    complexity param=0.2866826
  mean=0.2588235, MSE=0.1918339
  left son=2 (95 obs) right son=3 (75 obs)
  Primary splits:
      V6  < -0.2319769 to the right, improve=0.2807245, (0 missing)
      V14 < 0.806444   to the left,  improve=0.2088595, (0 missing)
      V18 < -0.8243789 to the right, improve=0.1836505, (0 missing)
      V3  < -0.4494587 to the right, improve=0.1435168, (0 missing)
      V11 < 1.747889   to the left,  improve=0.1414141, (0 missing)
  Surrogate splits:
      V3  < -0.5445668 to the right, agree=0.776, adj=0.493, (0 split)
      V10 < 0.03452701 to the right, agree=0.753, adj=0.440, (0 split)
      V18 < -0.6899486 to the right, agree=0.747, adj=0.427, (0 split)
      V12 < -0.5060276 to the right, agree=0.735, adj=0.400, (0 split)
      V1  < -0.6288789 to the right, agree=0.724, adj=0.373, (0 split)

Node number 2: 95 observations,    complexity param=0.06236768
  mean=0.05263158, MSE=0.0498615
  left son=4 (88 obs) right son=5 (7 obs)
  Primary splits:
      V18 < -0.6899486 to the right, improve=0.4293831, (0 missing)
      V3  < -0.6079721 to the right, improve=0.3214286, (0 missing)
      V14 < 0.405748   to the left,  improve=0.2847059, (0 missing)
      V1  < -0.6288789 to the right, improve=0.2540284, (0 missing)
      V17 < -0.8813544 to the right, improve=0.1916507, (0 missing)
  Surrogate splits:
```

```
      V4  < -1.148443  to the right, agree=0.968, adj=0.571, (0 split)
      V5  < -1.165512  to the right, agree=0.958, adj=0.429, (0 split)
      V14 < 0.6728787  to the left,  agree=0.958, adj=0.429, (0 split)
      V1  < -1.296803  to the right, agree=0.947, adj=0.286, (0 split)
      V3  < -1.083512  to the right, agree=0.947, adj=0.286, (0 split)

Node number 3: 75 observations,    complexity param=0.2866826
  mean=0.52, MSE=0.2496
  left son=6 (24 obs) right son=7 (51 obs)
  Primary splits:
      V12 < -0.8031538 to the left,  improve=0.5098039, (0 missing)
      V8  < 0.840574   to the right, improve=0.4791667, (0 missing)
      V11 < -0.8162264 to the left,  improve=0.4791667, (0 missing)
      V2  < -0.5448582 to the left,  improve=0.4617028, (0 missing)
      V7  < -0.7922774 to the left,  improve=0.4615385, (0 missing)
  Surrogate splits:
      V7  < -0.7922774 to the left,  agree=0.987, adj=0.958, (0 split)
      V8  < 0.7125586  to the right, agree=0.987, adj=0.958, (0 split)
      V11 < -0.8162264 to the left,  agree=0.987, adj=0.958, (0 split)
      V9  < -0.8034842 to the left,  agree=0.933, adj=0.792, (0 split)
      V3  < -1.337134  to the left,  agree=0.840, adj=0.500, (0 split)

Node number 4: 88 observations
  mean=0.01136364, MSE=0.0112345

Node number 5: 7 observations
  mean=0.5714286, MSE=0.244898

Node number 6: 24 observations
  mean=0, MSE=0

Node number 7: 51 observations,    complexity param=0.08236987
  mean=0.7647059, MSE=0.1799308
  left son=14 (10 obs) right son=15 (41 obs)
  Primary splits:
      V10 < -0.9299492 to the left,  improve=0.2927298, (0 missing)
      V2  < -0.7069363 to the left,  improve=0.2071417, (0 missing)
      V15 < 0.4316346  to the right, improve=0.1803002, (0 missing)
      V3  < -0.4494587 to the right, improve=0.1453297, (0 missing)
      V14 < -0.1285132 to the left,  improve=0.1453297, (0 missing)
  Surrogate splits:
      V2  < -1.031092  to the left,  agree=0.961, adj=0.8, (0 split)
      V13 < -1.066269  to the left,  agree=0.922, adj=0.6, (0 split)
      V16 < 0.6606819  to the right, agree=0.882, adj=0.4, (0 split)
      V18 < 0.7887851  to the right, agree=0.863, adj=0.3, (0 split)
      V14 < -1.06347   to the left,  agree=0.843, adj=0.2, (0 split)

Node number 14: 10 observations
  mean=0.3, MSE=0.21

Node number 15: 41 observations,    complexity param=0.03135378
  mean=0.8780488, MSE=0.1070791
  left son=30 (10 obs) right son=31 (31 obs)
  Primary splits:
      V15 < 0.2283145  to the right, improve=0.2329032, (0 missing)
      V3  < -0.4494587 to the right, improve=0.1960784, (0 missing)
      V7  < -0.4614003 to the right, improve=0.1608187, (0 missing)
```

```
      V11 < -0.2906623 to the right, improve=0.1608187, (0 missing)
      V12 < -0.5116872 to the right, improve=0.1608187, (0 missing)
  Surrogate splits:
      V1  < 1.678494   to the right, agree=0.805, adj=0.2, (0 split)
      V10 < 1.136785   to the right, agree=0.780, adj=0.1, (0 split)

Node number 30: 10 observations
  mean=0.6, MSE=0.24

Node number 31: 31 observations
  mean=0.9677419, MSE=0.03121748
```





In [24]:
```
bus_training_predict <- predict(fit,bus_train)
bus_predict <- predict(fit, bus_test)

tabel_mat_train <-table(bus_train$B,bus_training_predict)
table_mat <- table(bus_test$B, bus_predict)

accuracy_Train <- sum(diag(tabel_mat_train)) / sum(tabel_mat_train)
accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
print(paste('Accuracy for train', accuracy_Train))
print(paste('Accuracy for test', accuracy_Test))
```

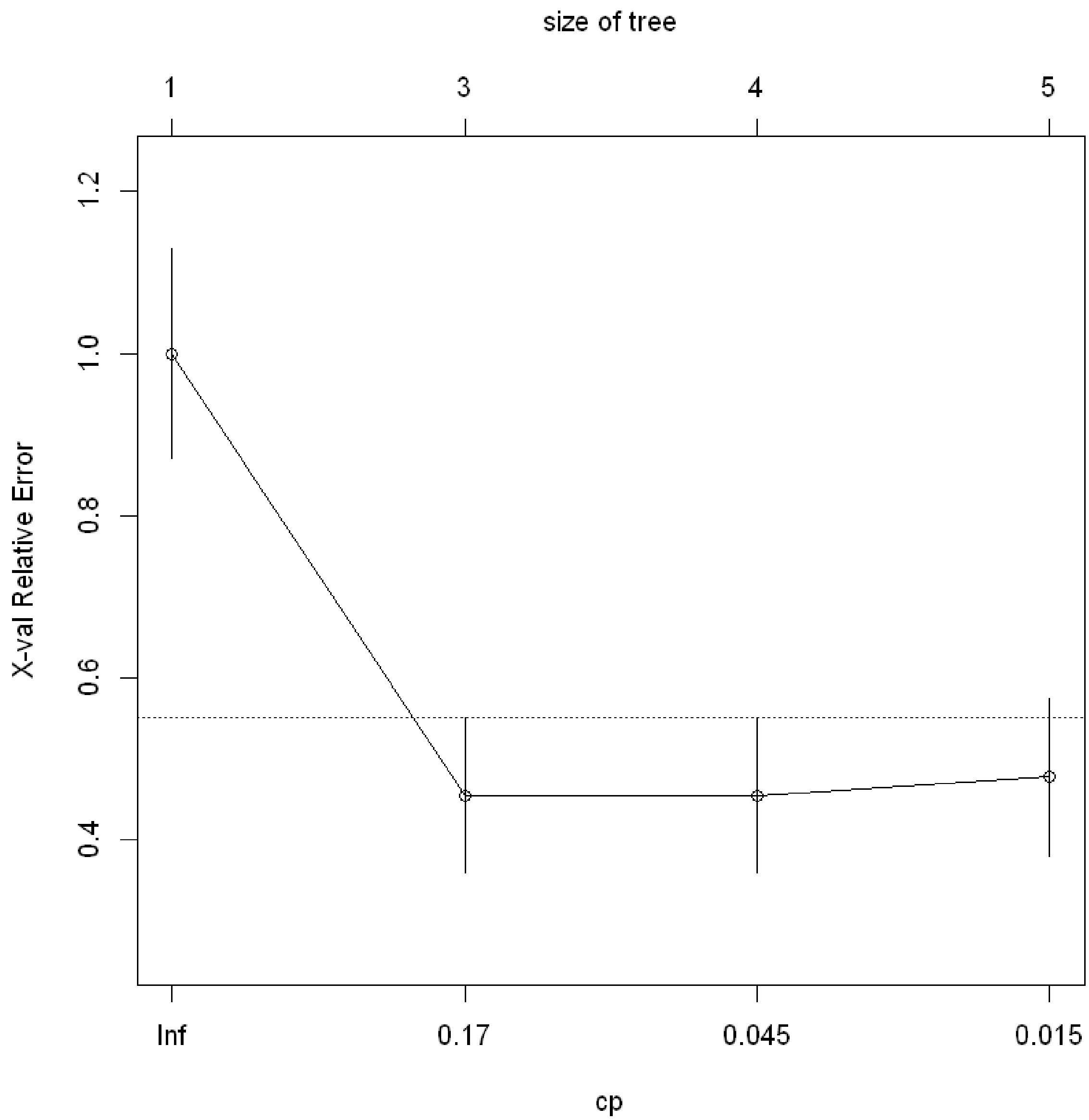
```
[1] "Accuracy for train 0.143491124260355"
[1] "Accuracy for test 0.147058823529412"
```

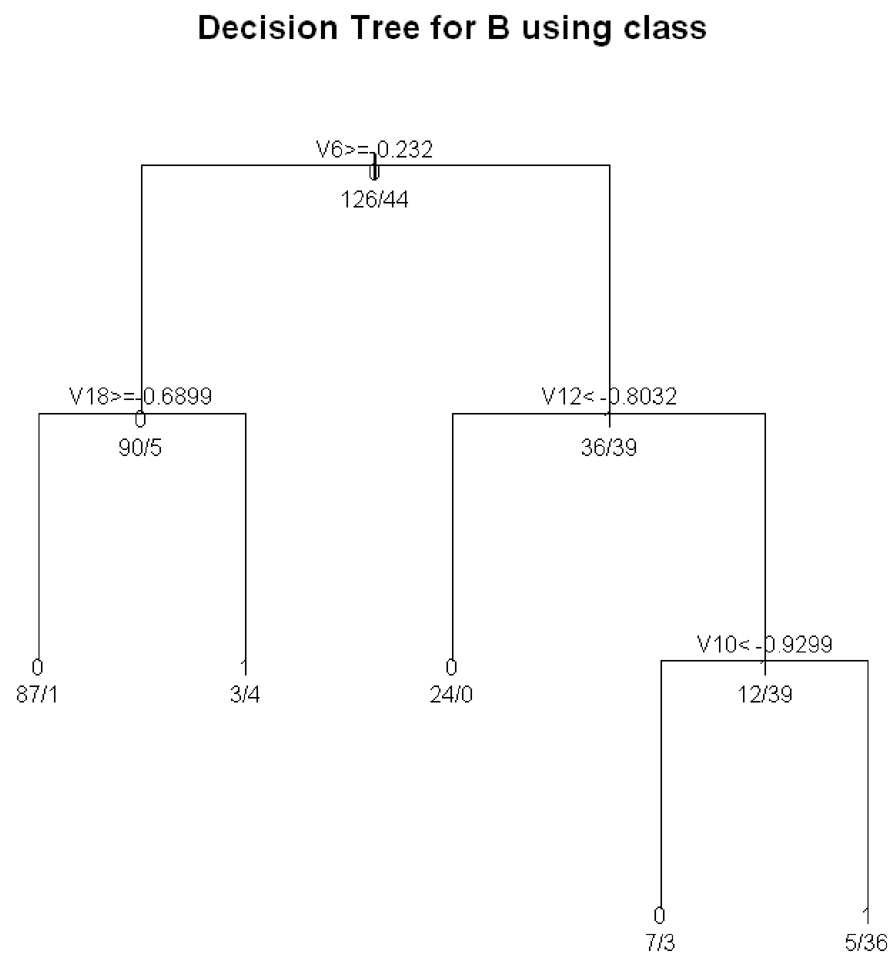
In [25]:
```
fit <-rpart(B~.,
            method="class", data=bus_test)
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits

plot(fit, uniform=TRUE,
     main="Decision Tree for B using class",margin=0.05)
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

```
Classification tree:
rpart(formula = B ~ ., data = bus_test, method = "class")

Variables actually used in tree construction:
[1] V10 V12 V18 V6

Root node error: 44/170 = 0.25882

n= 170

        CP nsplit rel error  xerror     xstd
1 0.306818      0   1.00000 1.00000 0.129788
2 0.090909      2   0.38636 0.45455 0.095474
3 0.022727      3   0.29545 0.45455 0.095474
4 0.010000      4   0.27273 0.47727 0.097505
Call:
rpart(formula = B ~ ., data = bus_test, method = "class")
  n= 170

          CP nsplit rel error    xerror       xstd
1 0.30681818      0 1.0000000 1.0000000 0.12978798
2 0.09090909      2 0.3863636 0.4545455 0.09547364
3 0.02272727      3 0.2954545 0.4545455 0.09547364
4 0.01000000      4 0.2727273 0.4772727 0.09750472

Variable importance
V12  V3  V6 V11  V7  V8  V9 V18 V10  V1  V2 V13 V14  V4 V16  V5
 14  11  10  10  10  10   8   7   7   4   2   2   2   1   1   1

Node number 1: 170 observations,    complexity param=0.3068182
  predicted class=0  expected loss=0.2588235  P(node) =1
    class counts:   126    44
   probabilities: 0.741 0.259
  left son=2 (95 obs) right son=3 (75 obs)
  Primary splits:
      V6  < -0.2319769 to the right, improve=18.309850, (0 missing)
      V14 < 0.806444   to the left,  improve=13.622560, (0 missing)
      V18 < -0.8243789 to the right, improve=11.978340, (0 missing)
      V3  < -0.4494587 to the right, improve= 9.360672, (0 missing)
      V8  < -1.463703  to the right, improve= 9.223529, (0 missing)
  Surrogate splits:
      V3  < -0.5445668 to the right, agree=0.776, adj=0.493, (0 split)
      V10 < 0.03452701 to the right, agree=0.753, adj=0.440, (0 split)
      V18 < -0.6899486 to the right, agree=0.747, adj=0.427, (0 split)
      V12 < -0.5060276 to the right, agree=0.735, adj=0.400, (0 split)
      V1  < -0.6288789 to the right, agree=0.724, adj=0.373, (0 split)

Node number 2: 95 observations,    complexity param=0.02272727
  predicted class=0  expected loss=0.05263158  P(node) =0.5588235
    class counts:    90     5
   probabilities: 0.947 0.053
  left son=4 (88 obs) right son=5 (7 obs)
  Primary splits:
      V18 < -0.6899486 to the right, improve=4.067840, (0 missing)
      V3  < -0.6079721 to the right, improve=3.045113, (0 missing)
      V14 < 0.405748   to the left,  improve=2.697214, (0 missing)
      V1  < -0.6288789 to the right, improve=2.406585, (0 missing)
```

```
        V17 < -0.8813544 to the right, improve=1.815638, (0 missing)
    Surrogate splits:
        V4  < -1.148443  to the right, agree=0.968, adj=0.571, (0 split)
        V5  < -1.165512  to the right, agree=0.958, adj=0.429, (0 split)
        V14 < 0.6728787  to the left,  agree=0.958, adj=0.429, (0 split)
        V1  < -1.296803  to the right, agree=0.947, adj=0.286, (0 split)
        V3  < -1.083512  to the right, agree=0.947, adj=0.286, (0 split)

Node number 3: 75 observations,    complexity param=0.3068182
  predicted class=1  expected loss=0.48  P(node) =0.4411765
    class counts:    36    39
   probabilities: 0.480 0.520 
  left son=6 (24 obs) right son=7 (51 obs)
  Primary splits:
      V12 < -0.8031538 to the left,  improve=19.08706, (0 missing)
      V8  < 0.840574   to the right, improve=17.94000, (0 missing)
      V11 < -0.8162264 to the left,  improve=17.94000, (0 missing)
      V2  < -0.5448582 to the left,  improve=17.28615, (0 missing)
      V7  < -0.7922774 to the left,  improve=17.28000, (0 missing)
  Surrogate splits:
      V7  < -0.7922774 to the left,  agree=0.987, adj=0.958, (0 split)
      V8  < 0.7125586  to the right, agree=0.987, adj=0.958, (0 split)
      V11 < -0.8162264 to the left,  agree=0.987, adj=0.958, (0 split)
      V9  < -0.8034842 to the left,  agree=0.933, adj=0.792, (0 split)
      V3  < -1.337134  to the left,  agree=0.840, adj=0.500, (0 split)

Node number 4: 88 observations
  predicted class=0  expected loss=0.01136364  P(node) =0.5176471
    class counts:    87     1
   probabilities: 0.989 0.011 

Node number 5: 7 observations
  predicted class=1  expected loss=0.4285714  P(node) =0.04117647
    class counts:     3     4
   probabilities: 0.429 0.571 

Node number 6: 24 observations
  predicted class=0  expected loss=0  P(node) =0.1411765
    class counts:    24     0
   probabilities: 1.000 0.000 

Node number 7: 51 observations,    complexity param=0.09090909
  predicted class=1  expected loss=0.2352941  P(node) =0.3
    class counts:    12    39
   probabilities: 0.235 0.765 
  left son=14 (10 obs) right son=15 (41 obs)
  Primary splits:
      V10 < -0.9299492 to the left,  improve=5.372453, (0 missing)
      V2  < -0.7069363 to the left,  improve=3.801659, (0 missing)
      V15 < 0.4316346  to the right, improve=3.309039, (0 missing)
      V3  < -0.4494587 to the right, improve=2.667227, (0 missing)
      V14 < -0.1285132 to the left,  improve=2.667227, (0 missing)
  Surrogate splits:
      V2  < -1.031092  to the left,  agree=0.961, adj=0.8, (0 split)
      V13 < -1.066269  to the left,  agree=0.922, adj=0.6, (0 split)
      V16 < 0.6606819  to the right, agree=0.882, adj=0.4, (0 split)
      V18 < 0.7887851  to the right, agree=0.863, adj=0.3, (0 split)
```

```
        V14 < -1.06347   to the left,  agree=0.843, adj=0.2, (0 split)

Node number 14: 10 observations
  predicted class=0  expected loss=0.3  P(node) =0.05882353
    class counts:     7     3
   probabilities: 0.700 0.300

Node number 15: 41 observations
  predicted class=1  expected loss=0.1219512  P(node) =0.2411765
    class counts:     5    36
   probabilities: 0.122 0.878
```





In [26]:
```
bus_training_predict <- predict(fit,bus_train, type="class")
bus_predict <- predict(fit, bus_test, type="class")

tabel_mat_train <-table(bus_train$B,bus_training_predict)
table_mat <- table(bus_test$B, bus_predict)

accuracy_Train <- sum(diag(tabel_mat_train)) / sum(tabel_mat_train)
accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
print(paste('Accuracy for train', accuracy_Train))
print(paste('Accuracy for test', accuracy_Test))
```

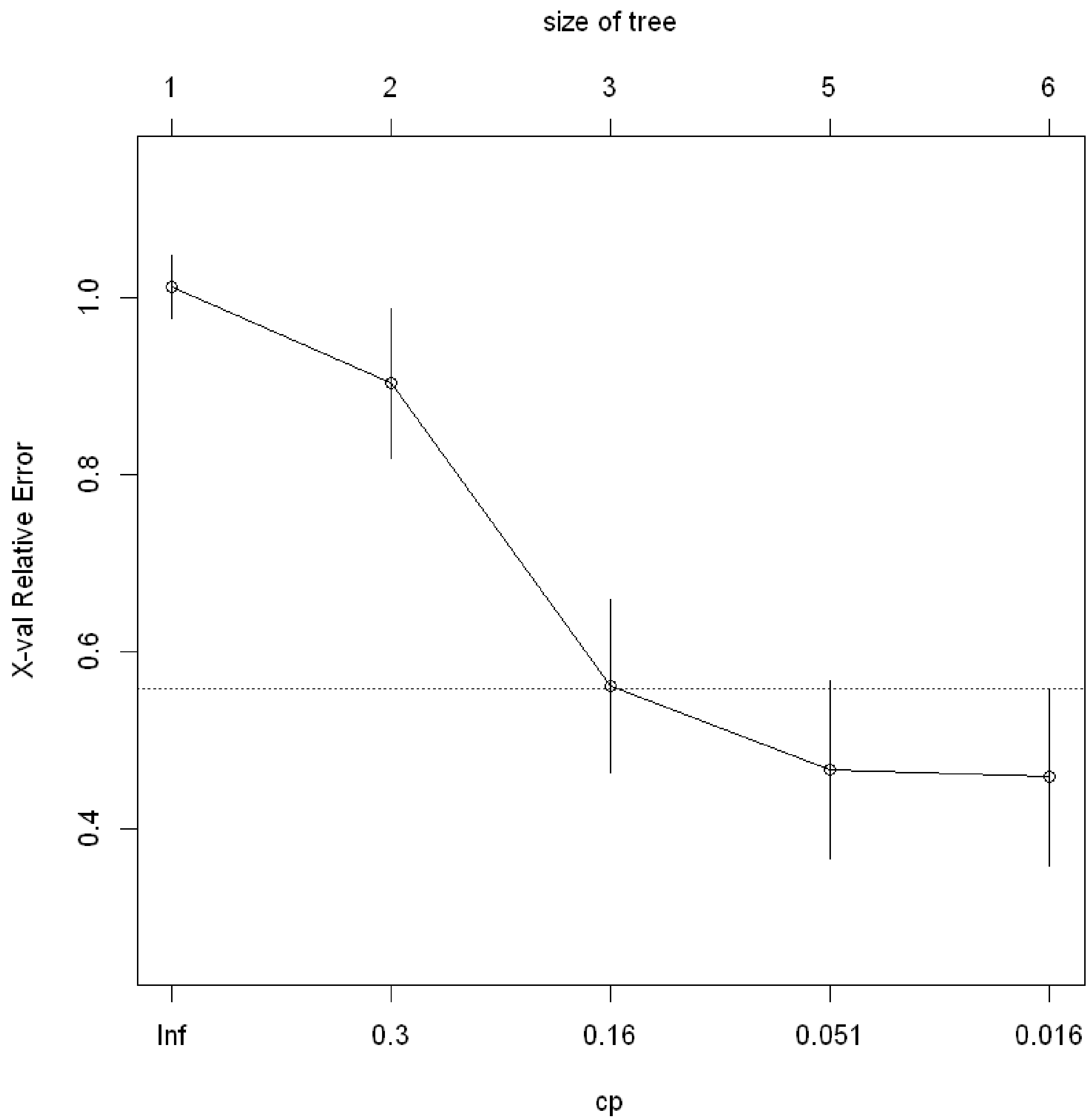
```
[1] "Accuracy for train 0.914201183431953"
[1] "Accuracy for test 0.929411764705882"
```

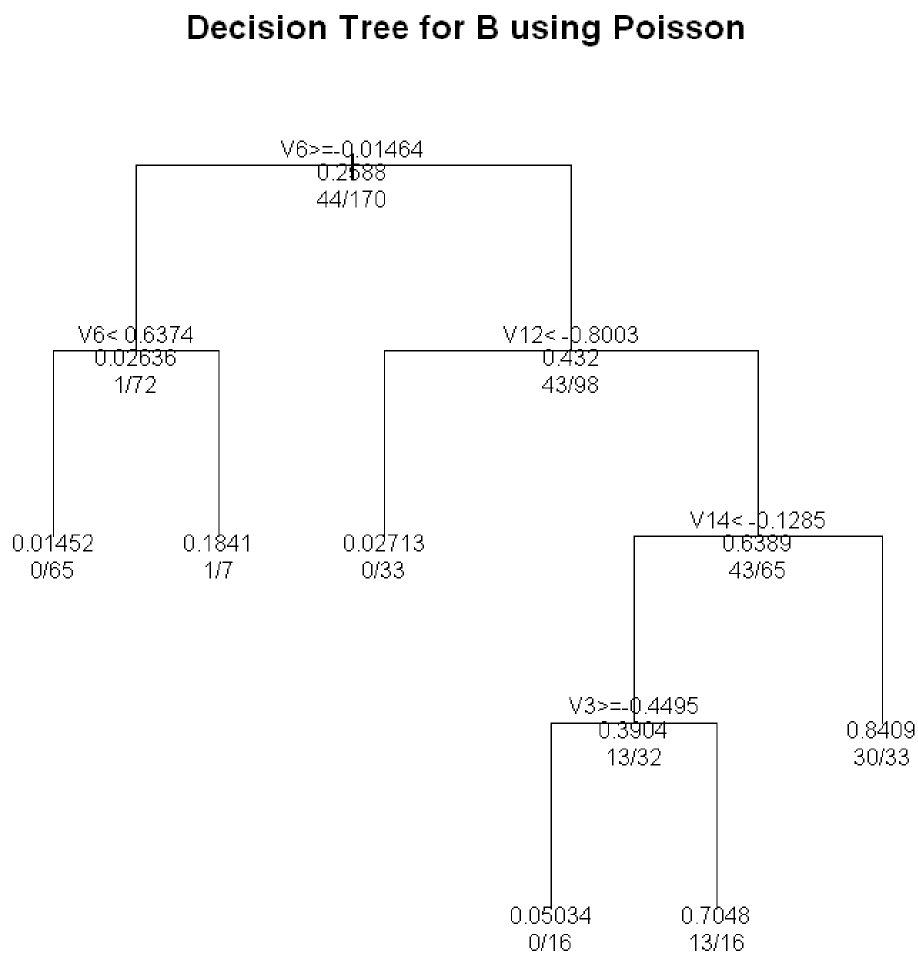
In [27]:
```
fit <-rpart(B~.,
            method="poisson", data=bus_test)
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits

plot(fit, uniform=TRUE,
     main="Decision Tree for B using Poisson",margin=0.05)
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

```
Rates regression tree:
rpart(formula = B ~ ., data = bus_test, method = "poisson")

Variables actually used in tree construction:
[1] V12 V14 V3  V6

Root node error: 118.94/170 = 0.69966

n= 170

        CP nsplit rel error  xerror     xstd
1 0.328054      0   1.00000 1.01189 0.035010
2 0.281472      1   0.67195 0.90294 0.084565
3 0.094285      2   0.39047 0.56029 0.097329
4 0.027056      4   0.20190 0.46672 0.100847
5 0.010000      5   0.17485 0.45811 0.099922
Call:
rpart(formula = B ~ ., data = bus_test, method = "poisson")
  n= 170

          CP nsplit rel error    xerror       xstd
1 0.32805351      0 1.0000000 1.0118879 0.03501034
2 0.28147178      1 0.6719465 0.9029362 0.08456515
3 0.09428514      2 0.3904747 0.5602883 0.09732926
4 0.02705634      4 0.2019044 0.4667222 0.10084701
5 0.01000000      5 0.1748481 0.4581074 0.09992200

Variable importance
 V7  V3  V9 V12  V6 V11  V8 V10  V2  V5 V14 V18 V17  V4  V1
 15  13  13  11  10  10   8   6   5   4   2   1   1   1   1

Node number 1: 170 observations,    complexity param=0.3280535
  events=44,  estimated rate=0.2588235 , mean deviance=0.6996563
  left son=2 (72 obs) right son=3 (98 obs)
  Primary splits:
      V6  < -0.01464306 to the right, improve=39.54425, (0 missing)
      V11 < -0.8003002  to the left,  improve=21.59954, (0 missing)
      V12 < -0.800324   to the left,  improve=20.94036, (0 missing)
      V14 < 0.806444    to the left,  improve=20.62782, (0 missing)
      V18 < -0.8243789  to the right, improve=19.63493, (0 missing)
  Surrogate splits:
      V10 < 0.03452701  to the right, agree=0.841, adj=0.625, (0 split)
      V3  < -0.03732373 to the right, agree=0.835, adj=0.611, (0 split)
      V2  < 0.2655322   to the right, agree=0.782, adj=0.486, (0 split)
      V7  < 0.03491524  to the right, agree=0.771, adj=0.458, (0 split)
      V9  < -0.4177024  to the right, agree=0.771, adj=0.458, (0 split)

Node number 2: 72 observations,    complexity param=0.02705634
  events=1,  estimated rate=0.02636309 , mean deviance=0.1259426
  left son=4 (65 obs) right son=5 (7 obs)
  Primary splits:
      V6  < 0.6373584   to the left,  improve=4.661512, (0 missing)
      V10 < -0.2754832  to the right, improve=4.661512, (0 missing)
      V3  < -0.3543506  to the right, improve=4.394449, (0 missing)
      V18 < -0.2866576  to the right, improve=4.394449, (0 missing)
      V1  < -0.5074382  to the right, improve=4.158883, (0 missing)
  Surrogate splits:
```

```
        V4 < 1.749486    to the left,  agree=0.931, adj=0.286, (0 split)
        V5 < 0.9895915   to the left,  agree=0.931, adj=0.286, (0 split)

Node number 3: 98 observations,    complexity param=0.2814718
  events=43,  estimated rate=0.43195 , mean deviance=0.7230052
  left son=6 (33 obs) right son=7 (65 obs)
  Primary splits:
      V12 < -0.800324   to the left,  improve=35.30990, (0 missing)
      V7  < -0.8524368  to the left,  improve=31.42954, (0 missing)
      V8  < 0.840574    to the right, improve=31.42954, (0 missing)
      V11 < -0.8003002  to the left,  improve=31.42954, (0 missing)
      V2  < -0.5448582  to the left,  improve=21.21730, (0 missing)
  Surrogate splits:
      V7  < -0.7922774  to the left,  agree=0.990, adj=0.970, (0 split)
      V8  < 0.7125586   to the right, agree=0.990, adj=0.970, (0 split)
      V11 < -0.8003002  to the left,  agree=0.969, adj=0.909, (0 split)
      V9  < -0.8034842  to the left,  agree=0.929, adj=0.788, (0 split)
      V3  < -0.9567017  to the left,  agree=0.796, adj=0.394, (0 split)

Node number 4: 65 observations
  events=0,  estimated rate=0.01452145 , mean deviance=0.0290429

Node number 5: 7 observations
  events=1,  estimated rate=0.1841004 , mean deviance=0.5659934

Node number 6: 33 observations
  events=0,  estimated rate=0.027127 , mean deviance=0.05425401

Node number 7: 65 observations,    complexity param=0.09428514
  events=43,  estimated rate=0.6389439 , mean deviance=0.5474681
  left son=14 (32 obs) right son=15 (33 obs)
  Primary splits:
      V14 < -0.1285132  to the left,  improve=6.395034, (0 missing)
      V10 < -0.8955036  to the left,  improve=5.746932, (0 missing)
      V3  < -0.5445668  to the right, improve=5.429323, (0 missing)
      V2  < -1.031092   to the left,  improve=4.703332, (0 missing)
      V6  < -0.2319769  to the right, improve=4.587401, (0 missing)
  Surrogate splits:
      V18 < -0.7571637  to the right, agree=0.908, adj=0.812, (0 split)
      V17 < -0.5568871  to the right, agree=0.892, adj=0.781, (0 split)
      V4  < -0.4911809  to the right, agree=0.754, adj=0.500, (0 split)
      V1  < -0.7503195  to the right, agree=0.723, adj=0.438, (0 split)
      V5  < -0.4682728  to the right, agree=0.708, adj=0.406, (0 split)

Node number 14: 32 observations,    complexity param=0.09428514
  events=13,  estimated rate=0.3903676 , mean deviance=0.7325267
  left son=28 (16 obs) right son=29 (16 obs)
  Primary splits:
      V3  < -0.4494587  to the right, improve=18.021830, (0 missing)
      V5  < 0.5458937   to the left,  improve=10.740090, (0 missing)
      V6  < -0.4493107  to the right, improve= 6.542553, (0 missing)
      V11 < -0.1314004  to the right, improve= 5.193881, (0 missing)
      V7  < -0.386201   to the right, improve= 4.903542, (0 missing)
  Surrogate splits:
      V5  < 0.5458937   to the left,  agree=0.844, adj=0.688, (0 split)
      V7  < -0.4614003  to the right, agree=0.781, adj=0.562, (0 split)
      V9  < -0.4177024  to the right, agree=0.781, adj=0.562, (0 split)
```

```
      V11 < -0.3065885  to the right, agree=0.781, adj=0.562, (0 split)
      V12 < -0.3617092  to the right, agree=0.781, adj=0.562, (0 split)

Node number 15: 33 observations
  events=30,  estimated rate=0.8409371 , mean deviance=0.1786713

Node number 28: 16 observations
  events=0,  estimated rate=0.05034325 , mean deviance=0.1006865

Node number 29: 16 observations
  events=13,  estimated rate=0.7048055 , mean deviance=0.3530903
```





In [28]:
```
bus_training_predict <- predict(fit,bus_train)
bus_predict <- predict(fit, bus_test)

tabel_mat_train <-table(bus_train$B,bus_training_predict)
table_mat <- table(bus_test$B, bus_predict)

accuracy_Train <- sum(diag(tabel_mat_train)) / sum(tabel_mat_train)
accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
print(paste('Accuracy for train', accuracy_Train))
print(paste('Accuracy for test', accuracy_Test))
```

```
[1] "Accuracy for train 0.368343195266272"
[1] "Accuracy for test 0.382352941176471"
```

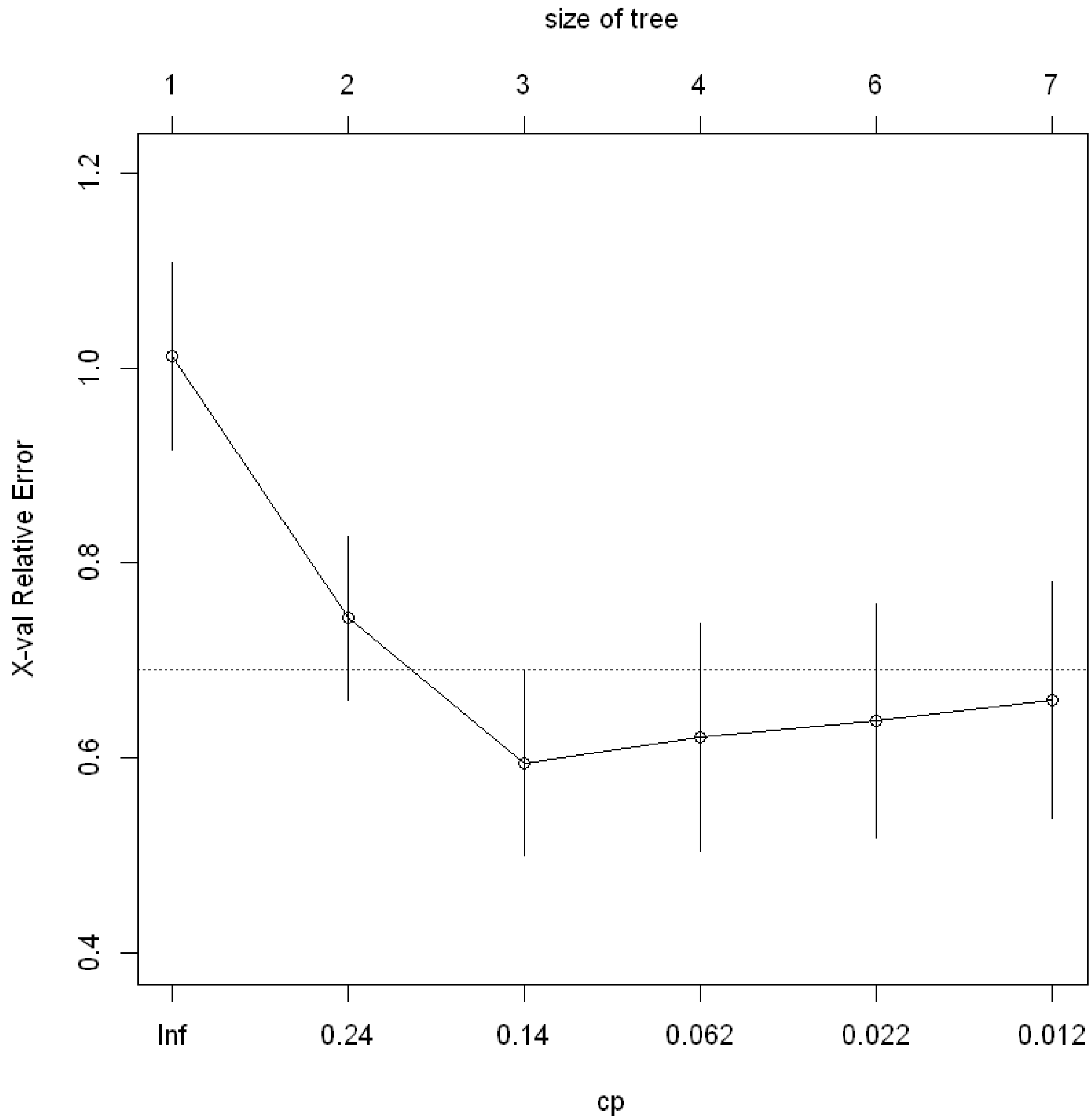
In [29]:
```
van_data <- select(zscore,V1,V2,V3,V4,V5,V6,V7,V8,V9,V10,V11,V12,V13,V14,V15,V
16,V17,V18, V)
set.seed(123)
sample <- sample.split(van_data$V, SplitRatio=0.8)

van_train = subset(van_data, sample == TRUE)
van_test = subset(van_data, sample == FALSE)
```

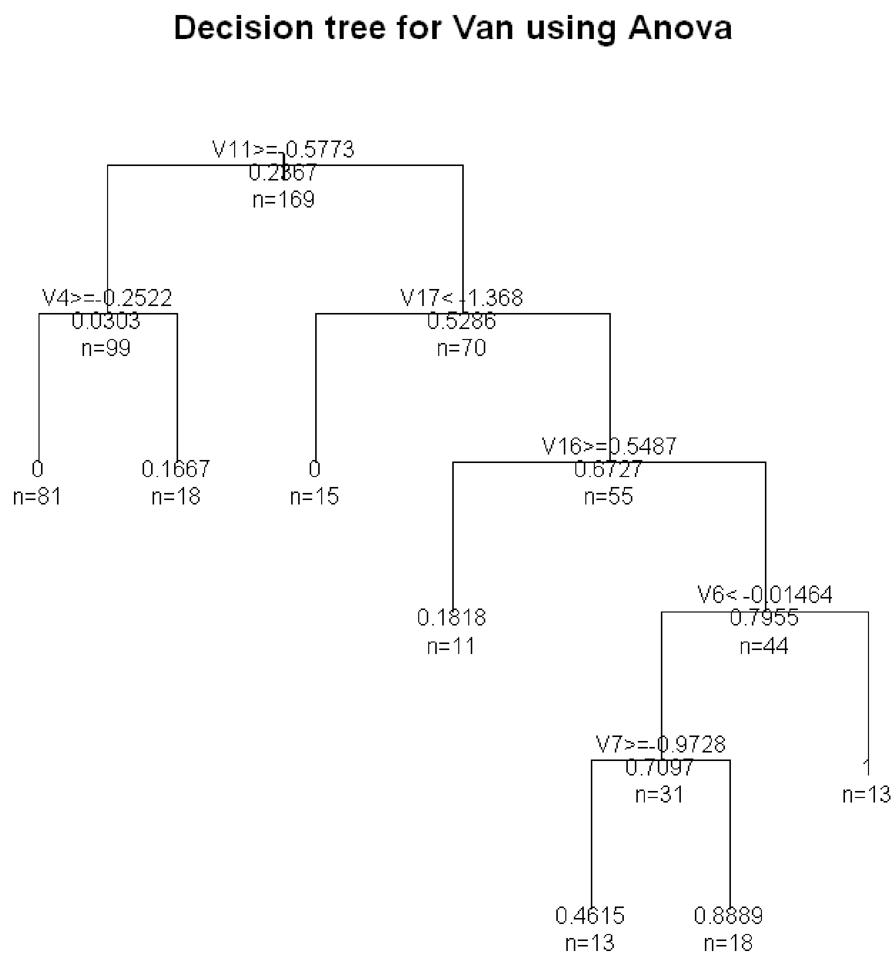
In [30]:
```
fit <-rpart(V~.,
            method="anova", data=van_test)
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits
plot(fit, uniform=TRUE,
     main="Decision tree for Van using Anova",margin=0.05)
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

```
Regression tree:
rpart(formula = V ~ ., data = van_test, method = "anova")

Variables actually used in tree construction:
[1] V11 V16 V17 V4  V6  V7 

Root node error: 30.533/169 = 0.18067

n= 169 

        CP nsplit rel error  xerror     xstd
1 0.333434      0   1.00000 1.01232 0.096516
2 0.174691      1   0.66657 0.74335 0.084245
3 0.108528      2   0.49187 0.59432 0.095242
4 0.035217      3   0.38335 0.62105 0.116721
5 0.013399      5   0.31291 0.63776 0.120068
6 0.010000      6   0.29951 0.65936 0.121273
Call:
rpart(formula = V ~ ., data = van_test, method = "anova")
  n= 169 

          CP nsplit rel error    xerror       xstd
1 0.33343426      0 1.0000000 1.0123189 0.09651629
2 0.17469118      1 0.6665657 0.7433502 0.08424520
3 0.10852801      2 0.4918746 0.5943191 0.09524204
4 0.03521724      3 0.3833465 0.6210524 0.11672078
5 0.01339852      5 0.3129121 0.6377587 0.12006769
6 0.01000000      6 0.2995135 0.6593567 0.12127324

Variable importance
V11 V12  V8  V7  V4  V9 V17 V14 V16 V18  V1 V10  V6 V13  V2 
 15  12  12  12  11   8   7   5   4   4   4   1   1   1   1 

Node number 1: 169 observations,    complexity param=0.3334343
  mean=0.2366864, MSE=0.1806659 
  left son=2 (99 obs) right son=3 (70 obs)
  Primary splits:
      V11 < -0.5773336  to the right, improve=0.3334343, (0 missing)
      V8  < 0.2004969   to the left,  improve=0.2721792, (0 missing)
      V12 < -0.3843474  to the right, improve=0.2301606, (0 missing)
      V4  < -0.222301   to the right, improve=0.2271689, (0 missing)
      V7  < -0.7321179  to the right, improve=0.2155511, (0 missing)
  Surrogate splits:
      V8  < 0.2004969   to the left,  agree=0.882, adj=0.714, (0 split)
      V12 < -0.6192186  to the right, agree=0.882, adj=0.714, (0 split)
      V7  < -0.5516395  to the right, agree=0.876, adj=0.700, (0 split)
      V4  < -0.4164921  to the right, agree=0.858, adj=0.657, (0 split)
      V9  < -0.8034842  to the right, agree=0.822, adj=0.571, (0 split)

Node number 2: 99 observations,    complexity param=0.01339852
  mean=0.03030303, MSE=0.02938476 
  left son=4 (81 obs) right son=5 (18 obs)
  Primary splits:
      V4  < -0.2521765  to the right, improve=0.14062500, (0 missing)
      V8  < 0.2004969   to the left,  improve=0.12343750, (0 missing)
      V17 < -0.5568871  to the right, improve=0.09765625, (0 missing)
      V11 < -0.2588099  to the right, improve=0.09250000, (0 missing)
```

```
        V12 < -0.3843474  to the right, improve=0.08333333, (0 missing)
    Surrogate splits:
        V1  < -0.7503195  to the right, agree=0.939, adj=0.667, (0 split)
        V8  < 0.3285124   to the left,  agree=0.929, adj=0.611, (0 split)
        V3  < -0.6079721  to the right, agree=0.919, adj=0.556, (0 split)
        V12 < -0.4437726  to the right, agree=0.919, adj=0.556, (0 split)
        V11 < -0.4499242  to the right, agree=0.899, adj=0.444, (0 split)

Node number 3: 70 observations,    complexity param=0.1746912
  mean=0.5285714, MSE=0.2491837
  left son=6 (15 obs) right son=7 (55 obs)
  Primary splits:
      V17 < -1.368055   to the left,  improve=0.3057851, (0 missing)
      V6  < -0.01464306 to the left,  improve=0.2432432, (0 missing)
      V14 < 0.9400093   to the right, improve=0.2335145, (0 missing)
      V1  < -1.236082   to the left,  improve=0.1867023, (0 missing)
      V18 < -1.496531   to the left,  improve=0.1867023, (0 missing)
  Surrogate splits:
      V14 < 0.9400093   to the right, agree=0.943, adj=0.733, (0 split)
      V18 < -1.496531   to the left,  agree=0.914, adj=0.600, (0 split)
      V1  < -1.357523   to the left,  agree=0.886, adj=0.467, (0 split)
      V4  < -1.357572   to the left,  agree=0.843, adj=0.267, (0 split)

Node number 4: 81 observations
  mean=0, MSE=0

Node number 5: 18 observations
  mean=0.1666667, MSE=0.1388889

Node number 6: 15 observations
  mean=0, MSE=0

Node number 7: 55 observations,    complexity param=0.108528
  mean=0.6727273, MSE=0.2201653
  left son=14 (11 obs) right son=15 (44 obs)
  Primary splits:
      V16 < 0.5487154   to the right, improve=0.2736486, (0 missing)
      V10 < -0.3099288  to the left,  improve=0.2193980, (0 missing)
      V6  < -0.01464306 to the left,  improve=0.1824324, (0 missing)
      V2  < -1.355249   to the left,  improve=0.1084835, (0 missing)
      V7  < -0.4614003  to the left,  improve=0.1081081, (0 missing)
  Surrogate splits:
      V10 < -1.54997    to the left,  agree=0.818, adj=0.091, (0 split)
      V11 < -0.609186   to the right, agree=0.818, adj=0.091, (0 split)

Node number 14: 11 observations
  mean=0.1818182, MSE=0.1487603

Node number 15: 44 observations,    complexity param=0.03521724
  mean=0.7954545, MSE=0.1627066
  left son=30 (31 obs) right son=31 (13 obs)
  Primary splits:
      V6  < -0.01464306 to the left,  improve=0.10783410, (0 missing)
      V10 < -0.3099288  to the left,  improve=0.10779620, (0 missing)
      V4  < -0.9841276  to the right, improve=0.08217312, (0 missing)
      V7  < -0.5967591  to the left,  improve=0.07563025, (0 missing)
      V12 < -0.6192186  to the left,  improve=0.07563025, (0 missing)
```

```
  Surrogate splits:
      V7  < -0.5967591  to the left,  agree=0.932, adj=0.769, (0 split)
      V12 < -0.6192186  to the left,  agree=0.932, adj=0.769, (0 split)
      V8  < 0.4565278   to the right, agree=0.909, adj=0.692, (0 split)
      V10 < 0.03452701  to the left,  agree=0.909, adj=0.692, (0 split)
      V9  < -0.8034842  to the left,  agree=0.886, adj=0.615, (0 split)

Node number 30: 31 observations,    complexity param=0.03521724
  mean=0.7096774, MSE=0.2060354
  left son=60 (13 obs) right son=61 (18 obs)
  Primary splits:
      V7  < -0.9727557  to the right, improve=0.2158335, (0 missing)
      V12 < -0.9220043  to the right, improve=0.2158335, (0 missing)
      V8  < 0.9685894   to the left,  improve=0.1737603, (0 missing)
      V13 < -0.7590161  to the right, improve=0.1414983, (0 missing)
      V4  < -0.8496876  to the right, improve=0.1347687, (0 missing)
  Surrogate splits:
      V12 < -0.9220043  to the right, agree=1.000, adj=1.000, (0 split)
      V8  < 0.9685894   to the left,  agree=0.935, adj=0.846, (0 split)
      V11 < -0.9436359  to the right, agree=0.871, adj=0.692, (0 split)
      V13 < -0.2213237  to the right, agree=0.806, adj=0.538, (0 split)
      V2  < -0.8690144  to the right, agree=0.774, adj=0.462, (0 split)

Node number 31: 13 observations
  mean=1, MSE=0

Node number 60: 13 observations
  mean=0.4615385, MSE=0.2485207

Node number 61: 18 observations
  mean=0.8888889, MSE=0.09876543
```


size of tree
1
2
3
4
6
7
X-val Relative Error
0.4
0.6
0.8
1.0
1.2
Inf
0.24
0.14
0.062
0.022
0.012
cp



In [31]:

```
van_training_predict <- predict(fit,van_train)
van_predict <- predict(fit, van_test)

tabel_mat_train <-table(van_train$V,van_training_predict)
table_mat <- table(van_test$V, van_predict)

accuracy_Train <- sum(diag(tabel_mat_train)) / sum(tabel_mat_train)
accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
print(paste('Accuracy for train', accuracy_Train))
print(paste('Accuracy for test', accuracy_Test))
```

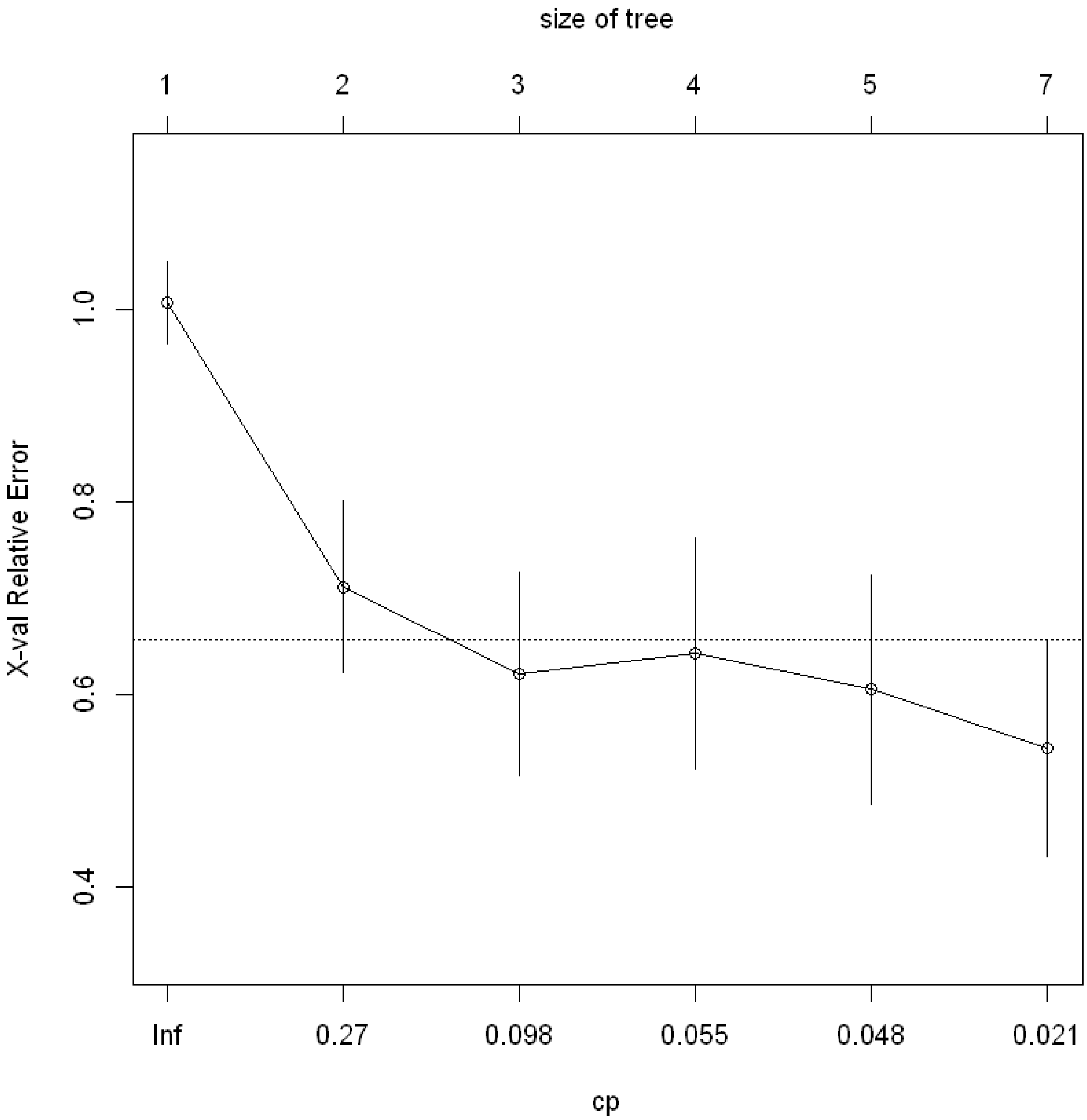
```
[1] "Accuracy for train 0.608567208271787"
[1] "Accuracy for test 0.585798816568047"
```

In [32]:
```
fit <-rpart(V~.,
            method="poisson", data=van_test)
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits
plot(fit, uniform=TRUE,
     main="Decision tree for Van using Poisson",margin=0.05)
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

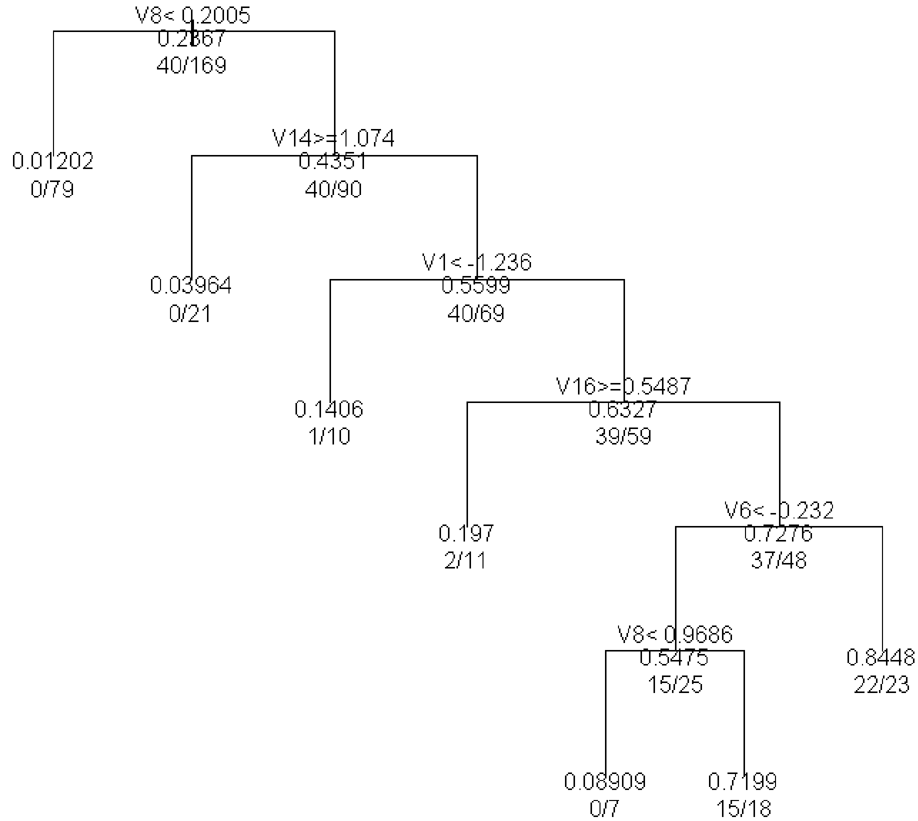
```
Rates regression tree:
rpart(formula = V ~ ., data = van_test, method = "poisson")

Variables actually used in tree construction:
[1] V1  V14 V16 V6  V8

Root node error: 115.28/169 = 0.68214

n= 169

        CP nsplit rel error  xerror     xstd
1 0.420630      0   1.00000 1.00667 0.043063
2 0.169683      1   0.57937 0.71200 0.089058
3 0.056958      2   0.40969 0.62105 0.105755
4 0.053347      3   0.35273 0.64299 0.119267
5 0.043708      4   0.29938 0.60537 0.118670
6 0.010000      6   0.21197 0.54430 0.112637
Call:
rpart(formula = V ~ ., data = van_test, method = "poisson")
  n= 169

          CP nsplit rel error    xerror       xstd
1 0.42062963      0 1.0000000 1.0066725 0.04306341
2 0.16968326      1 0.5793704 0.7120036 0.08905848
3 0.05695812      2 0.4096871 0.6210480 0.10575510
4 0.05334726      3 0.3527290 0.6429871 0.11926711
5 0.04370757      4 0.2993817 0.6053660 0.11867006
6 0.01000000      6 0.2119666 0.5443049 0.11263739

Variable importance
 V8 V12  V4  V7  V9 V11 V14 V18 V17  V1  V3 V16 V10
 15  15  14  13  12  12   5   4   4   3   2   2   1

Node number 1: 169 observations,    complexity param=0.4206296
  events=40,  estimated rate=0.2366864 , mean deviance=0.6821393
  left son=2 (79 obs) right son=3 (90 obs)
  Primary splits:
      V8  < 0.2004969   to the left,  improve=50.40712, (0 missing)
      V11 < -0.4499242  to the right, improve=48.38343, (0 missing)
      V12 < -0.3843474  to the right, improve=44.41502, (0 missing)
      V4  < 0.04657903  to the right, improve=42.78231, (0 missing)
      V7  < -0.2508422  to the right, improve=41.18226, (0 missing)
  Surrogate splits:
      V12 < -0.4607513  to the right, agree=0.976, adj=0.949, (0 split)
      V11 < -0.2588099  to the right, agree=0.970, adj=0.937, (0 split)
      V7  < -0.2508422  to the right, agree=0.935, adj=0.861, (0 split)
      V4  < -0.1774876  to the right, agree=0.923, adj=0.835, (0 split)
      V9  < -0.4177024  to the right, agree=0.905, adj=0.797, (0 split)

Node number 2: 79 observations
  events=0,  estimated rate=0.01201562 , mean deviance=0.02403124

Node number 3: 90 observations,    complexity param=0.1696833
  events=40,  estimated rate=0.4351287 , mean deviance=0.7210249
  left son=6 (21 obs) right son=7 (69 obs)
  Primary splits:
      V14 < 1.073575    to the right, improve=21.25625, (0 missing)
```

```
        V17 < -1.368055   to the left,  improve=21.25625, (0 missing)
        V1  < -1.357523   to the left,  improve=13.52611, (0 missing)
        V18 < -1.496531   to the left,  improve=13.25813, (0 missing)
        V6  < -0.01464306 to the left,  improve=12.70698, (0 missing)
    Surrogate splits:
        V18 < -1.496531   to the left,  agree=0.944, adj=0.762, (0 split)
        V17 < -1.205822   to the left,  agree=0.933, adj=0.714, (0 split)
        V4  < -1.357572   to the left,  agree=0.833, adj=0.286, (0 split)
        V1  < -1.478964   to the left,  agree=0.811, adj=0.190, (0 split)

Node number 6: 21 observations
  events=0,  estimated rate=0.03964321 , mean deviance=0.07928642

Node number 7: 69 observations,    complexity param=0.05695812
  events=40,  estimated rate=0.5599181 , mean deviance=0.6328388
  left son=14 (10 obs) right son=15 (59 obs)
  Primary splits:
      V1  < -1.236082   to the left,  improve=6.722882, (0 missing)
      V16 < 0.5487154   to the right, improve=6.660775, (0 missing)
      V6  < -0.01464306 to the left,  improve=6.623704, (0 missing)
      V10 < -1.343296   to the left,  improve=5.622647, (0 missing)
      V7  < -0.4614003  to the left,  improve=4.224364, (0 missing)

Node number 14: 10 observations
  events=1,  estimated rate=0.1405975 , mean deviance=0.4735658

Node number 15: 59 observations,    complexity param=0.05334726
  events=39,  estimated rate=0.6326611 , mean deviance=0.5485424
  left son=30 (11 obs) right son=31 (48 obs)
  Primary splits:
      V16 < 0.5487154   to the right, improve=6.210171, (0 missing)
      V10 < -1.343296   to the left,  improve=5.769029, (0 missing)
      V6  < -0.01464306 to the left,  improve=4.190028, (0 missing)
      V4  < -0.2372387  to the right, improve=2.868578, (0 missing)
      V2  < -1.355249   to the left,  improve=2.865246, (0 missing)
  Surrogate splits:
      V10 < -1.54997    to the left,  agree=0.864, adj=0.273, (0 split)

Node number 30: 11 observations
  events=2,  estimated rate=0.1970443 , mean deviance=0.6211165

Node number 31: 48 observations,    complexity param=0.04370757
  events=37,  estimated rate=0.7276209 , mean deviance=0.4037868
  left son=62 (25 obs) right son=63 (23 obs)
  Primary splits:
      V6  < -0.2319769  to the left,  improve=1.980303, (0 missing)
      V17 < 1.389917    to the right, improve=1.747938, (0 missing)
      V4  < -0.2372387  to the right, improve=1.446777, (0 missing)
      V11 < -0.5454813  to the right, improve=1.446777, (0 missing)
      V18 < 1.057646    to the right, improve=1.331061, (0 missing)
  Surrogate splits:
      V3  < -0.417756   to the left,  agree=0.833, adj=0.652, (0 split)
      V8  < 0.9685894   to the right, agree=0.812, adj=0.609, (0 split)
      V10 < -0.1721465  to the left,  agree=0.812, adj=0.609, (0 split)
      V12 < -0.9078554  to the left,  agree=0.792, adj=0.565, (0 split)
      V2  < -0.5448582  to the left,  agree=0.771, adj=0.522, (0 split)
```

```
Node number 62: 25 observations,    complexity param=0.04370757
  events=15,  estimated rate=0.5474765 , mean deviance=0.6178759
  left son=124 (7 obs) right son=125 (18 obs)
  Primary splits:
      V8  < 0.9685894   to the left,  improve=9.855122, (0 missing)
      V12 < -0.8767279  to the right, improve=9.855122, (0 missing)
      V7  < -0.9727557  to the right, improve=7.191441, (0 missing)
      V4  < -0.7003098  to the right, improve=4.396144, (0 missing)
      V3  < -1.242026   to the right, improve=2.902524, (0 missing)
  Surrogate splits:
      V12 < -0.8767279  to the right, agree=1.00, adj=1.000, (0 split)
      V7  < -0.942676   to the right, agree=0.96, adj=0.857, (0 split)
      V4  < -0.4911809  to the right, agree=0.92, adj=0.714, (0 split)
      V9  < -0.8034842  to the right, agree=0.92, adj=0.714, (0 split)
      V3  < -0.7347829  to the right, agree=0.88, adj=0.571, (0 split)

Node number 63: 23 observations
  events=22,  estimated rate=0.8448118 , mean deviance=0.09919846

Node number 124: 7 observations
  events=0,  estimated rate=0.08908686 , mean deviance=0.1781737

Node number 125: 18 observations
  events=15,  estimated rate=0.71991 , mean deviance=0.3208685
```


size of tree
1
2
3
4
5
7
X-val Relative Error
0.4
0.6
0.8
1.0
Inf
0.27
0.098
0.055
0.048
0.021
cp

Decision tree for Van using Poisson



In [33]:
```
van_training_predict <- predict(fit,van_train)
van_predict <- predict(fit, van_test)

tabel_mat_train <-table(van_train$V,van_training_predict)
table_mat <- table(van_test$V, van_predict)

accuracy_Train <- sum(diag(tabel_mat_train)) / sum(tabel_mat_train)
accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
print(paste('Accuracy for train', accuracy_Train))
print(paste('Accuracy for test', accuracy_Test))
```

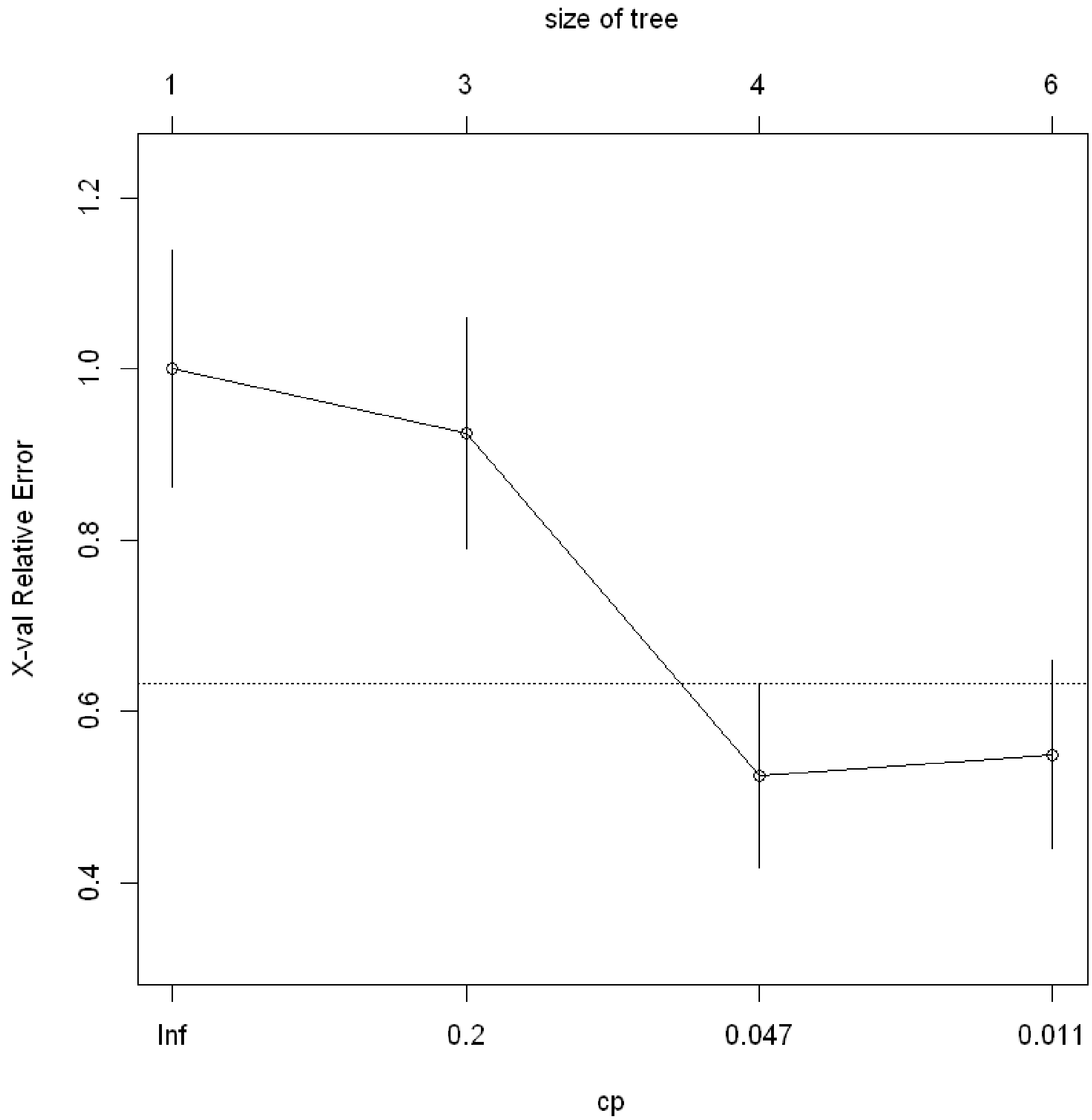
```
[1] "Accuracy for train 0.511078286558346"
[1] "Accuracy for test 0.467455621301775"
```

In [34]:
```
fit <-rpart(V~.,
            method="class", data=van_test)
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits
plot(fit, uniform=TRUE,
     main="Decision tree for Van using class",margin=0.05)
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

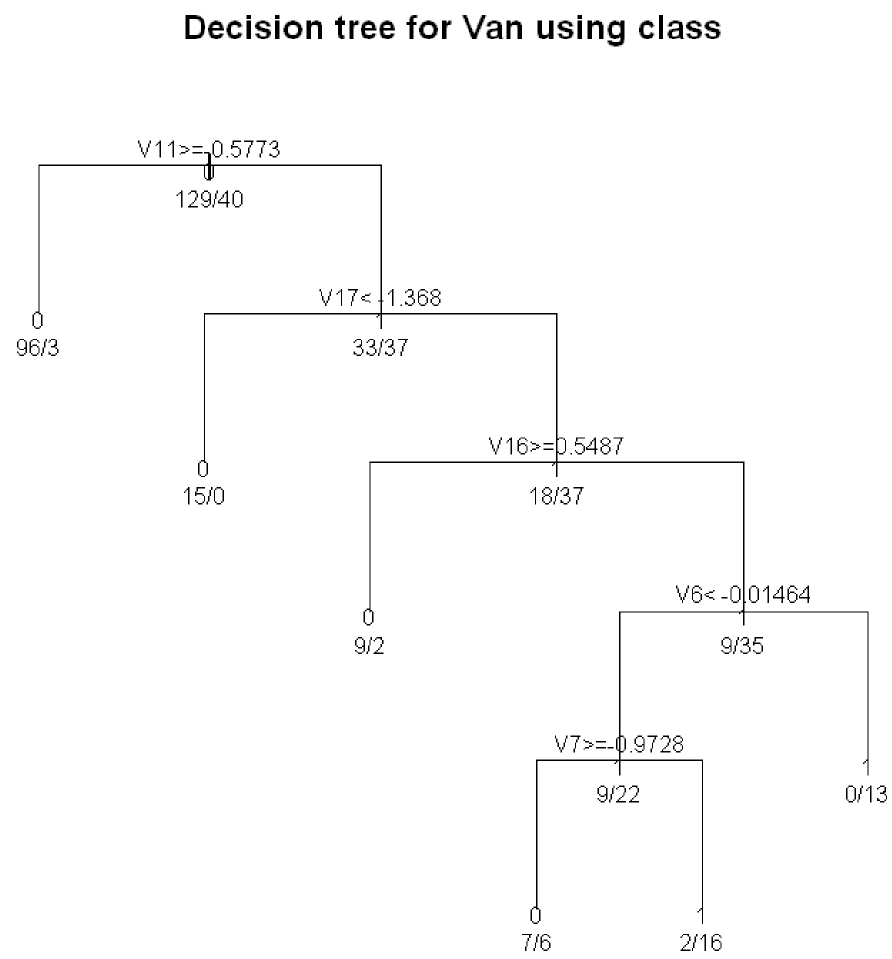
```
Classification tree:
rpart(formula = V ~ ., data = van_test, method = "class")

Variables actually used in tree construction:
[1] V11 V16 V17 V6  V7

Root node error: 40/169 = 0.23669

n= 169

      CP nsplit rel error xerror    xstd
1 0.2375      0     1.000  1.000 0.13814
2 0.1750      2     0.525  0.925 0.13440
3 0.0125      3     0.350  0.525 0.10721
4 0.0100      5     0.325  0.550 0.10936
Call:
rpart(formula = V ~ ., data = van_test, method = "class")
  n= 169

      CP nsplit rel error xerror      xstd
1 0.2375      0     1.000  1.000 0.1381407
2 0.1750      2     0.525  0.925 0.1343954
3 0.0125      3     0.350  0.525 0.1072105
4 0.0100      5     0.325  0.550 0.1093621

Variable importance
V11 V12  V7  V8  V4  V9 V17 V14 V16 V18  V1 V10  V6 V13  V2
 15  12  12  12  11   8   7   5   4   4   3   1   1   1   1

Node number 1: 169 observations,    complexity param=0.2375
  predicted class=0  expected loss=0.2366864  P(node) =1
    class counts:   129    40
   probabilities: 0.763 0.237
  left son=2 (99 obs) right son=3 (70 obs)
  Primary splits:
      V11 < -0.5773336  to the right, improve=20.36119, (0 missing)
      V8  < 0.2004969   to the left,  improve=16.62064, (0 missing)
      V12 < -0.3843474  to the right, improve=14.05478, (0 missing)
      V4  < -0.222301   to the right, improve=13.87209, (0 missing)
      V7  < -0.7321179  to the right, improve=13.16265, (0 missing)
  Surrogate splits:
      V8  < 0.2004969   to the left,  agree=0.882, adj=0.714, (0 split)
      V12 < -0.6192186  to the right, agree=0.882, adj=0.714, (0 split)
      V7  < -0.5516395  to the right, agree=0.876, adj=0.700, (0 split)
      V4  < -0.4164921  to the right, agree=0.858, adj=0.657, (0 split)
      V9  < -0.8034842  to the right, agree=0.822, adj=0.571, (0 split)

Node number 2: 99 observations
  predicted class=0  expected loss=0.03030303  P(node) =0.5857988
    class counts:    96     3
   probabilities: 0.970 0.030

Node number 3: 70 observations,    complexity param=0.2375
  predicted class=1  expected loss=0.4714286  P(node) =0.4142012
    class counts:    33    37
   probabilities: 0.471 0.529
  left son=6 (15 obs) right son=7 (55 obs)
```

```
    Primary splits:
        V17 < -1.368055   to the left,  improve=10.667530, (0 missing)
        V6  < -0.01464306 to the left,  improve= 8.485714, (0 missing)
        V14 < 0.9400093   to the right, improve= 8.146320, (0 missing)
        V1  < -1.236082   to the left,  improve= 6.513245, (0 missing)
        V18 < -1.496531   to the left,  improve= 6.513245, (0 missing)
    Surrogate splits:
        V14 < 0.9400093   to the right, agree=0.943, adj=0.733, (0 split)
        V18 < -1.496531   to the left,  agree=0.914, adj=0.600, (0 split)
        V1  < -1.357523   to the left,  agree=0.886, adj=0.467, (0 split)
        V4  < -1.357572   to the left,  agree=0.843, adj=0.267, (0 split)

Node number 6: 15 observations
  predicted class=0  expected loss=0  P(node) =0.0887574
    class counts:    15     0
   probabilities: 1.000 0.000 

Node number 7: 55 observations,    complexity param=0.175
  predicted class=1  expected loss=0.3272727  P(node) =0.3254438
    class counts:    18    37
   probabilities: 0.327 0.673 
  left son=14 (11 obs) right son=15 (44 obs)
  Primary splits:
      V16 < 0.5487154   to the right, improve=6.627273, (0 missing)
      V10 < -0.3099288  to the left,  improve=5.313420, (0 missing)
      V6  < -0.01464306 to the left,  improve=4.418182, (0 missing)
      V2  < -1.355249   to the left,  improve=2.627273, (0 missing)
      V7  < -0.4614003  to the left,  improve=2.618182, (0 missing)
  Surrogate splits:
      V10 < -1.54997    to the left,  agree=0.818, adj=0.091, (0 split)
      V11 < -0.609186   to the right, agree=0.818, adj=0.091, (0 split)

Node number 14: 11 observations
  predicted class=0  expected loss=0.1818182  P(node) =0.06508876
    class counts:     9     2
   probabilities: 0.818 0.182 

Node number 15: 44 observations,    complexity param=0.0125
  predicted class=1  expected loss=0.2045455  P(node) =0.260355
    class counts:     9    35
   probabilities: 0.205 0.795 
  left son=30 (31 obs) right son=31 (13 obs)
  Primary splits:
      V6  < -0.01464306 to the left,  improve=1.543988, (0 missing)
      V10 < -0.3099288  to the left,  improve=1.543445, (0 missing)
      V4  < -0.9841276  to the right, improve=1.176570, (0 missing)
      V7  < -0.5967591  to the left,  improve=1.082888, (0 missing)
      V12 < -0.6192186  to the left,  improve=1.082888, (0 missing)
  Surrogate splits:
      V7  < -0.5967591  to the left,  agree=0.932, adj=0.769, (0 split)
      V12 < -0.6192186  to the left,  agree=0.932, adj=0.769, (0 split)
      V8  < 0.4565278   to the right, agree=0.909, adj=0.692, (0 split)
      V10 < 0.03452701  to the left,  agree=0.909, adj=0.692, (0 split)
      V9  < -0.8034842  to the left,  agree=0.886, adj=0.615, (0 split)

Node number 30: 31 observations,    complexity param=0.0125
  predicted class=1  expected loss=0.2903226  P(node) =0.183432
```

```
      class counts:     9    22
     probabilities: 0.290 0.710
    left son=60 (13 obs) right son=61 (18 obs)
    Primary splits:
        V7  < -0.9727557  to the right, improve=2.757100, (0 missing)
        V12 < -0.9220043  to the right, improve=2.757100, (0 missing)
        V8  < 0.9685894   to the left,  improve=2.219648, (0 missing)
        V13 < -0.7590161  to the right, improve=1.807527, (0 missing)
        V4  < -0.8496876  to the right, improve=1.721562, (0 missing)
    Surrogate splits:
        V12 < -0.9220043  to the right, agree=1.000, adj=1.000, (0 split)
        V8  < 0.9685894   to the left,  agree=0.935, adj=0.846, (0 split)
        V11 < -0.9436359  to the right, agree=0.871, adj=0.692, (0 split)
        V13 < -0.2213237  to the right, agree=0.806, adj=0.538, (0 split)
        V2  < -0.8690144  to the right, agree=0.774, adj=0.462, (0 split)

Node number 31: 13 observations
  predicted class=1  expected loss=0  P(node) =0.07692308
    class counts:     0    13
   probabilities: 0.000 1.000

Node number 60: 13 observations
  predicted class=0  expected loss=0.4615385  P(node) =0.07692308
    class counts:     7     6
   probabilities: 0.538 0.462

Node number 61: 18 observations
  predicted class=1  expected loss=0.1111111  P(node) =0.1065089
    class counts:     2    16
   probabilities: 0.111 0.889
```



size of tree
1
3
4
6
X-val Relative Error
1.2
1.0
0.8
0.6
0.4
Inf
0.2
0.047
0.011
cp



In [35]:
```
van_training_predict <- predict(fit,van_train, type="class")
van_predict <- predict(fit, van_test, type="class")

tabel_mat_train <-table(van_train$V,van_training_predict)
table_mat <- table(van_test$V, van_predict)

accuracy_Train <- sum(diag(tabel_mat_train)) / sum(tabel_mat_train)
accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
print(paste('Accuracy for train', accuracy_Train))
print(paste('Accuracy for test', accuracy_Test))
```

```
[1] "Accuracy for train 0.846381093057607"
[1] "Accuracy for test 0.923076923076923"
```

Analysis:

The first step is to separate bus data from the zscore data into test and training set. The split ratio is 80/20, 80% of the data goes to the training set and the 20% goes to the test set. Next step is to train the bus data in the training set, for this we use the Decision tree regression algorithm. There are different methods for the Decision tree, the three different methods that were used was ANOVA, class, and Poisson. After training, we get a model that can be used to predict the test set using the predict method. Next step is to look at the accuracy of the three different models. These are the accuracy:

Bus Train Result

- Anova method: 0.143491124260355
- Class method: 0.914201183431953
- Poisson method: 0.368343195266272

Bus Test Result

- Anova method: 0.147058823529412
- Class method: 0.929411764705882
- Poisson method: 0.382352941176471

From these result for Bus, the Class method model has better accuracy.

The above step can be applied to predict for Van

Van Train Result

- Anova method: 0.608567208271787
- Class method: 0.846381093057607
- Poisson method: 0.511078286558346

Van Test Result

- Anova method: 0.585798816568047
- Class method: 0.923076923076923
- Poisson method: 0.467455621301775

From these result for Van, the Class method model has better accuracy.

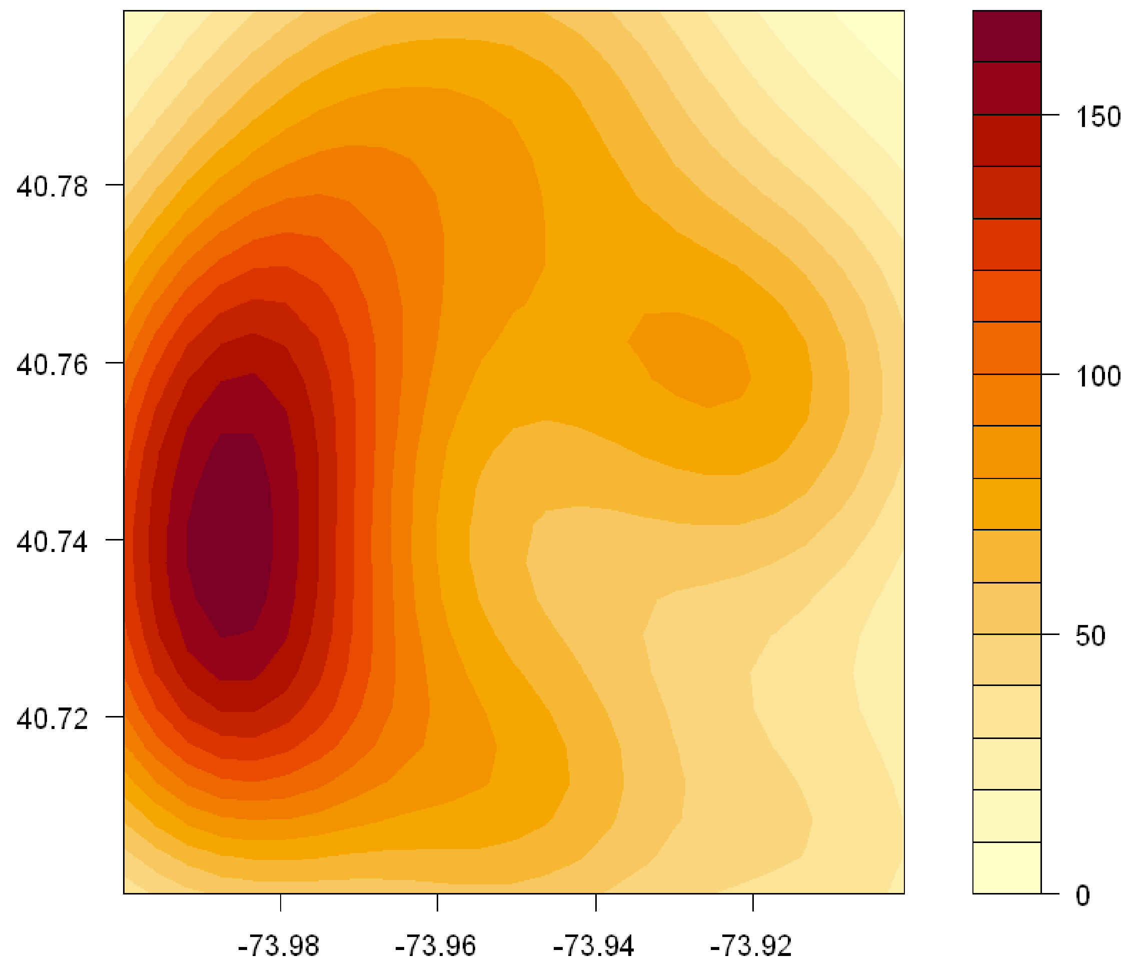
Number 8

The correlation of the data is very important as it impacts the visualization of everything ranging from question 2 to 5. The data has a more defined pattern/trend when the correlation between the two attributes that we are using has a high correlation. The pattern between circularity and hallow ratio is defined/similar due to its high correlation between each other. Additionally, for the decision tree, we found out using the class method gives you the best accuracy of the decision tree. When it comes to difficulty separating the data and understanding what each attribute means and how it contributes to the dependent variable.

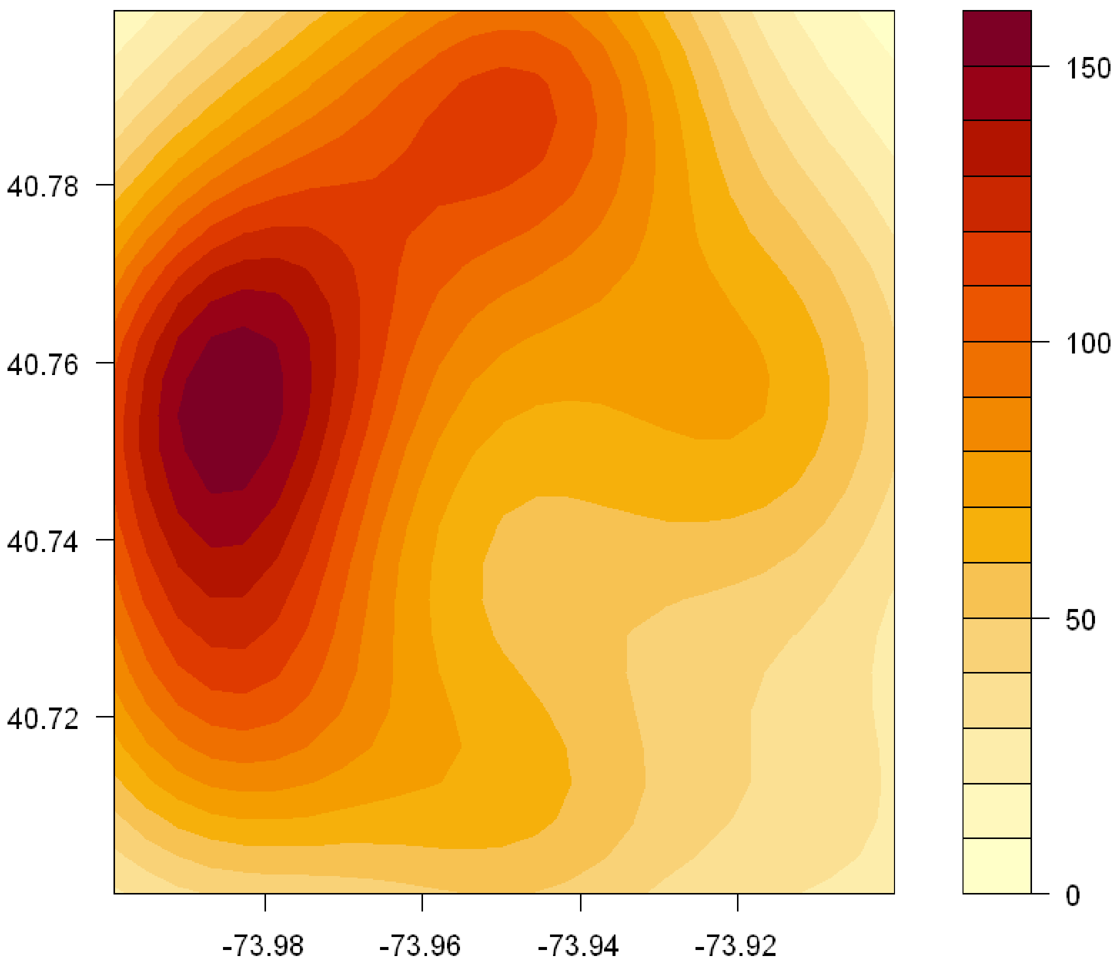
Part B

Number 10

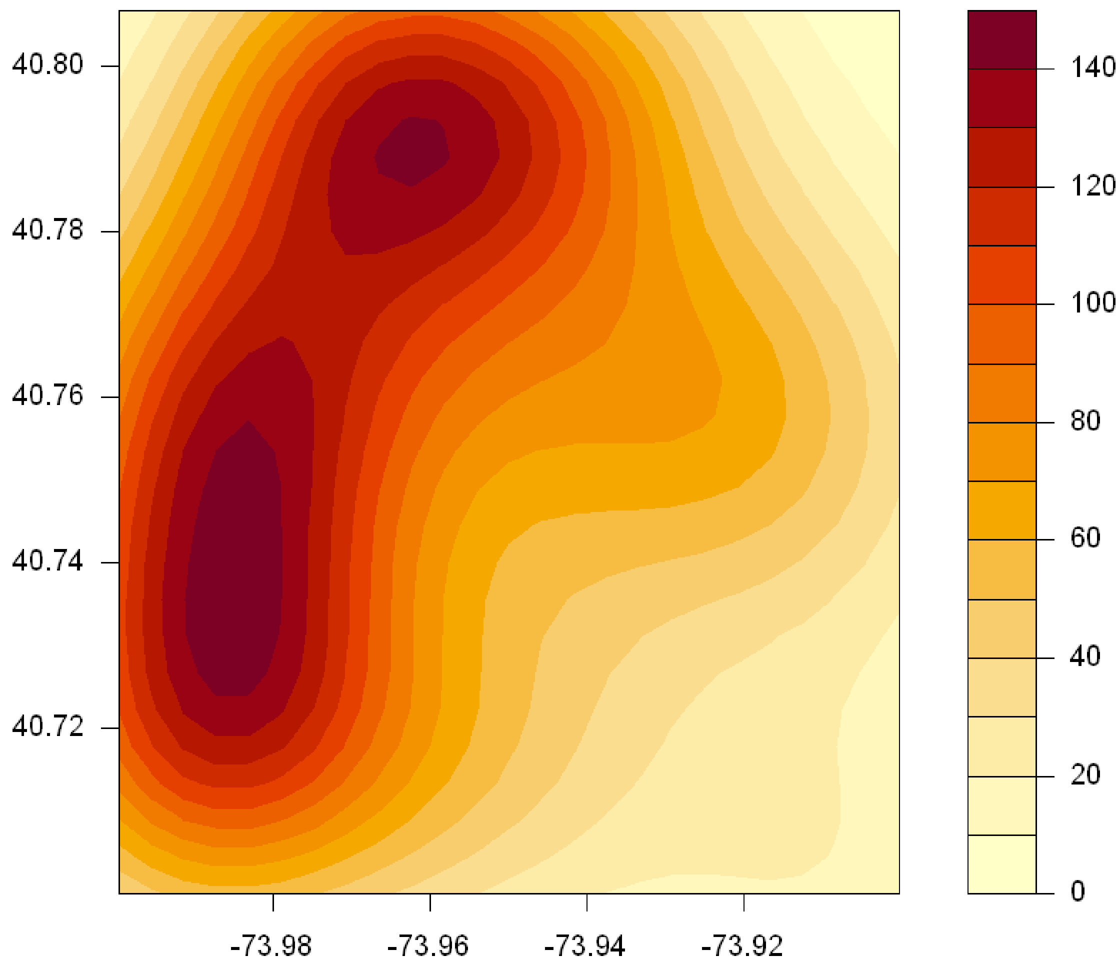
In [36]:
```
Harassment05 <- read.csv(file="Harassment0-5.csv", header=TRUE, sep=",")
k <- with(Harassment05,MASS:::kde2d(Harassment05$Longitude,Harassment05$Latitu
de, h = c(0.06, 0.06)))
filled.contour(k)
```



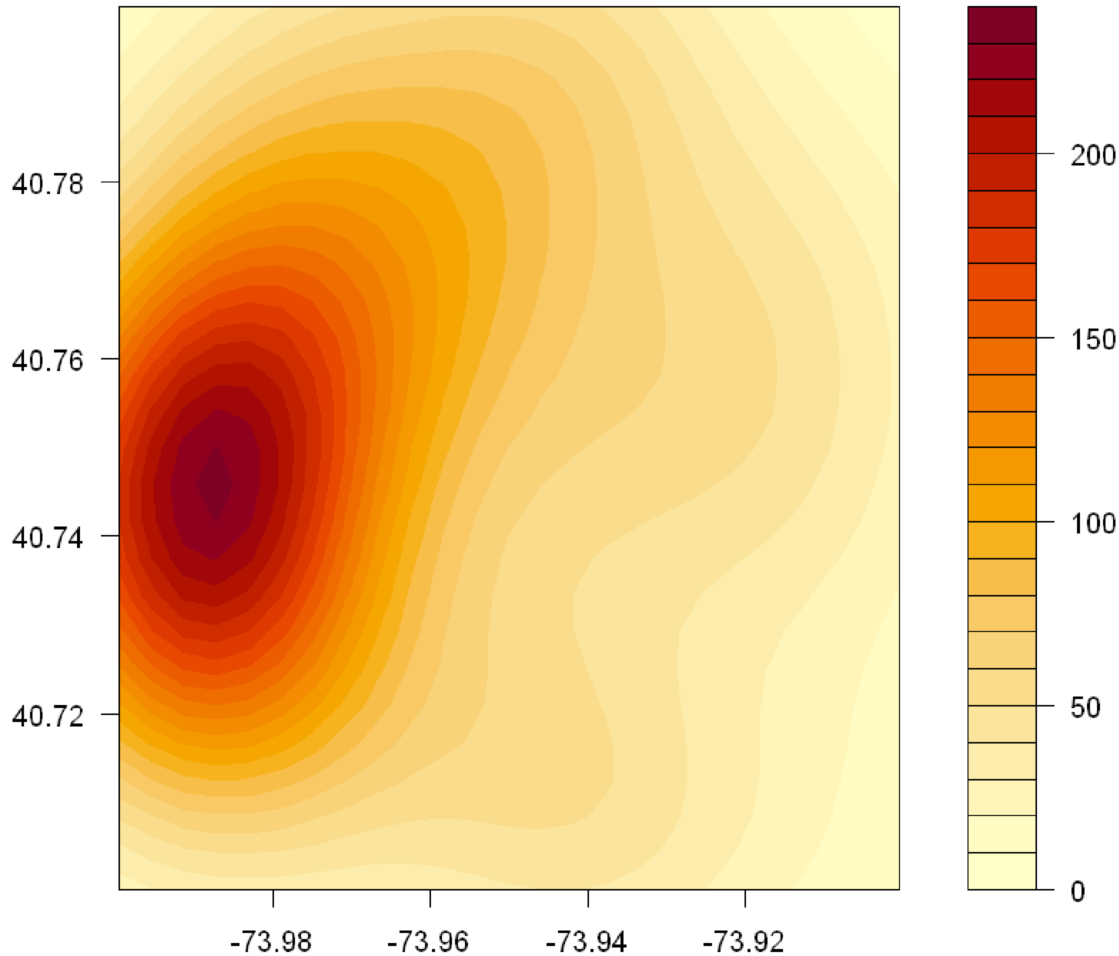
In [37]:
```
Harassment611 <- read.csv(file="Harassment6-11.csv", header=TRUE, sep=",")
k <- with(Harassment611,MASS:::kde2d(Harassment611$Longitude,Harassment611$Latitude, h = c(0.06, 0.06)))
filled.contour(k)
```



In [38]:
```
Harassment1217 <- read.csv(file="Harassment12-17.csv", header=TRUE, sep=",")
k <- with(Harassment1217,MASS:::kde2d(Harassment1217$Longitude,Harassment1217$
Latitude, h = c(0.06, 0.06)))
filled.contour(k)
```



In [39]:
```
PetitLarcency611 <- read.csv(file="PetitLarcency6-11.csv", header=TRUE, sep=
",")
k <- with(PetitLarcency611,MASS:::kde2d(PetitLarcency611$Longitude,PetitLarcen
cy611$Latitude, h = c(0.06, 0.06)))
filled.contour(k)
```

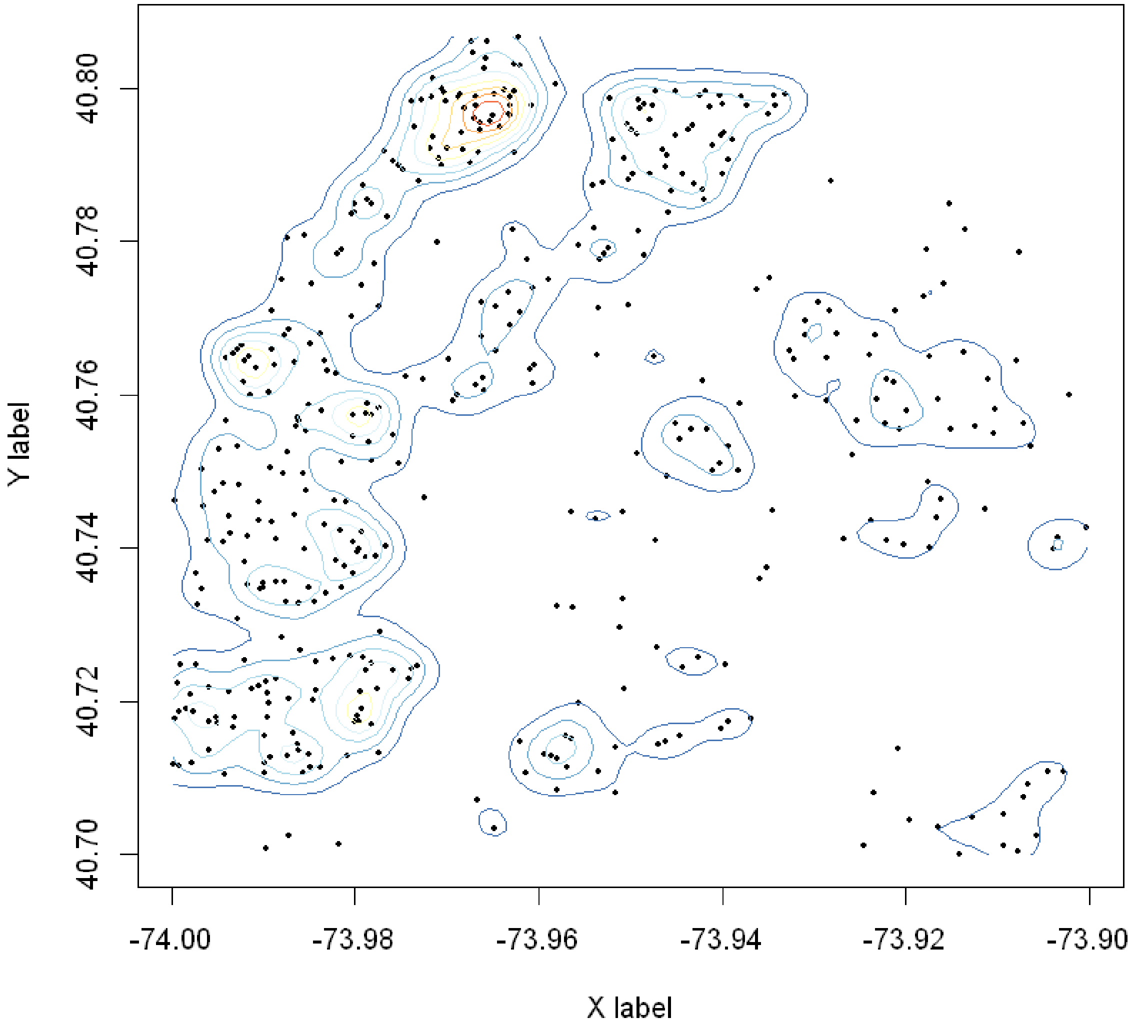


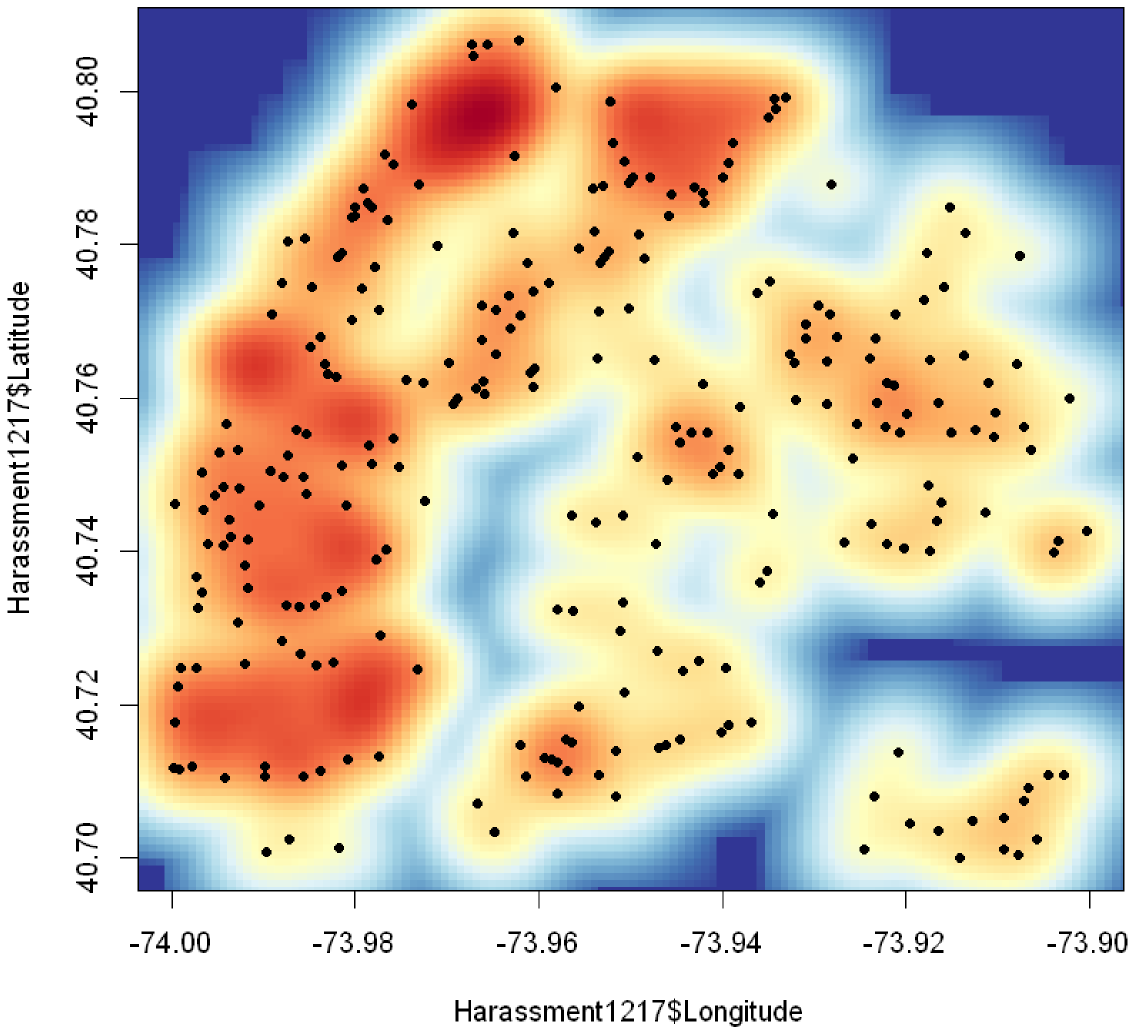
Analysis:

- For this question, I chose the vector bandwidths of c(0.06, 0.06). This vector bandwidths best fit the series of heatmap I created because the contour lines generated using the bandwidth aligns very well with the data represented on the heatmap.
- This bandwidth created a contour map that is concentrated around the high-volume area on the heat map, but it is ignoring the plots on the map. I incremented the bandwidth by 0.01, this showed improvements in the contour line be align to the points.

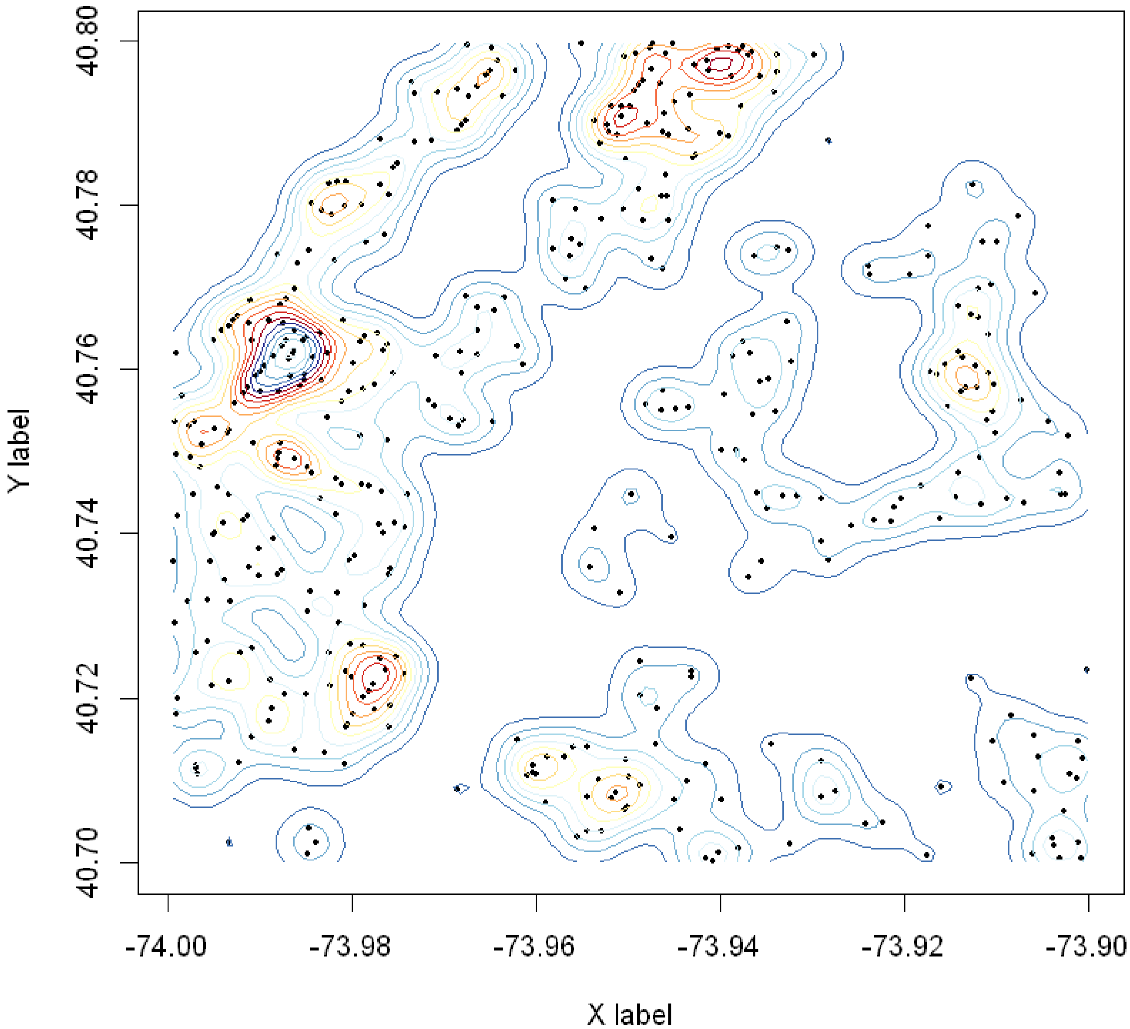
Number 11

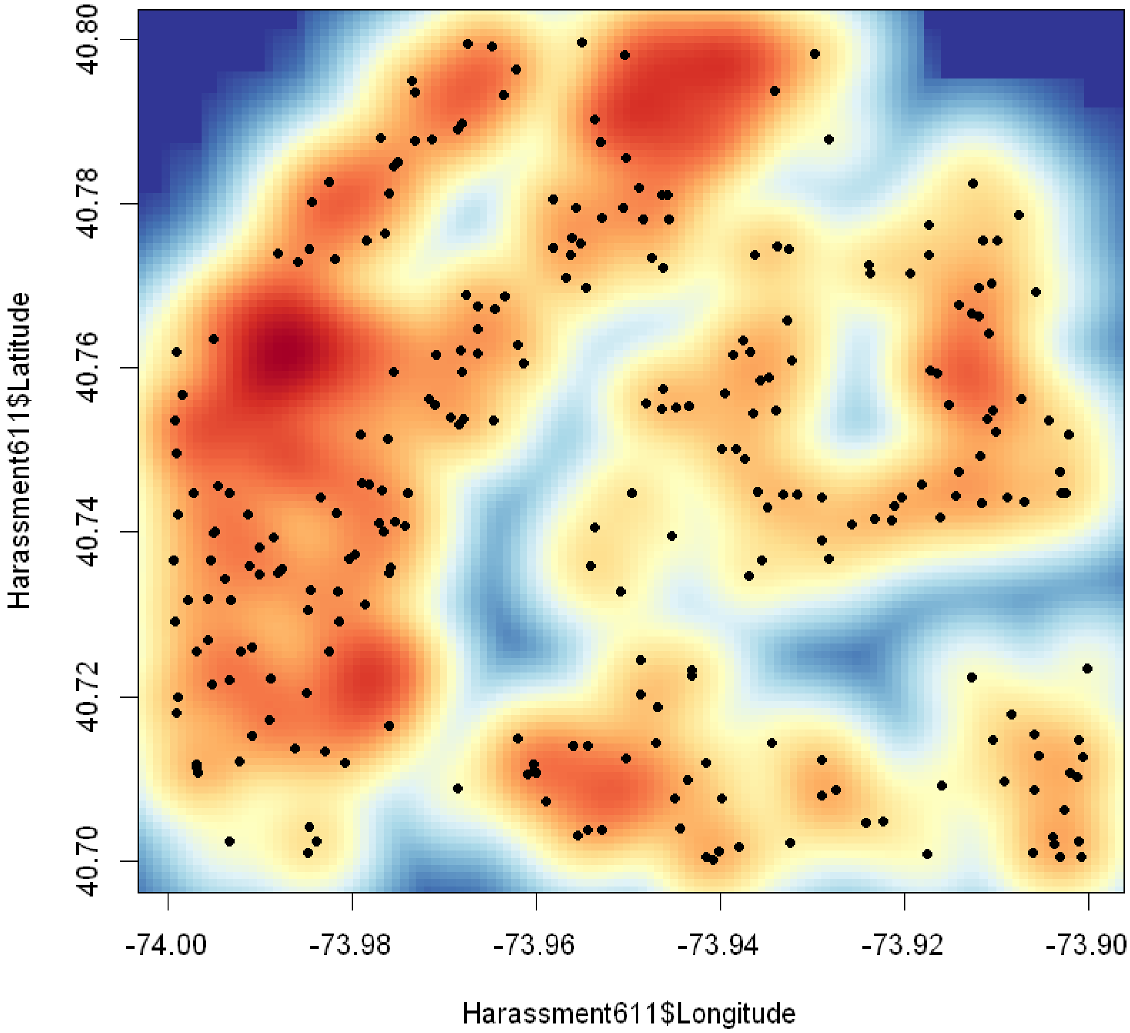
In [43]:
```
k <- 11
my.cols <- rev(brewer.pal(k, "RdYlBu"))
z <- kde2d(Harassment1217$Longitude,Harassment1217$Latitude , n=100,h = c(0.01
, 0.01))

plot(Harassment1217$Longitude,Harassment1217$Latitude, xlab="X label", ylab="Y
label", pch=19, cex=.4)
contour(z, drawlabels=FALSE, nlevels=k, col=my.cols, add=TRUE)
smoothScatter(Harassment1217$Longitude,Harassment1217$Latitude, nrpoints=.3*n,
colramp=colorRampPalette(my.cols), pch=19, cex=.8)
```


40.70
40.72
40.74
40.76
40.78
40.80
Y label
-74.00
-73.98
-73.96
-73.94
-73.92
-73.90
X label


40.70
40.72
40.74
40.76
40.78
40.80
Harassment1217$Latitude
-74.00
-73.98
-73.96
-73.94
-73.92
-73.90
Harassment1217$Longitude

In [46]:
```
k <- 11
my.cols <- rev(brewer.pal(k, "RdYlBu"))
z <- kde2d(Harassment611$Longitude,Harassment611$Latitude , n=100,h = c(0.01,
0.01))

plot(Harassment611$Longitude,Harassment611$Latitude, xlab="X label", ylab="Y l
abel", pch=19, cex=.4)
contour(z, drawlabels=FALSE, nlevels=k, col=my.cols, add=TRUE)
smoothScatter(Harassment611$Longitude,Harassment611$Latitude, nrpoints=.3*n, c
olramp=colorRampPalette(my.cols), pch=19, cex=.8)
```


40.80
40.78
40.76
40.74
40.72
40.70
Y label
-74.00
-73.98
-73.96
-73.94
-73.92
-73.90
X label



Analysis:

These density plots represent the crime rate based on the location with the x-axis being the longitude and the y-axis being the latitude. The darker color on the color spectrum means that more crimes or harassments are being committed in that area. Around ~(-73.9, 40.76) for Harassment6-11 the area is darker and the contour lines are closer, meaning there are more crime being commited.

Number 12

Analysis

The highest crime rate area for harassment and petit larceny in 6-11, that is collocated is between the longitude range of -74 to -73.98, and the latitude range of 40.74 to 40.76. The darker spots in those coordinates for both of the heatmaps. The anti collocation for these heatmaps identifies the locations in which crime does not occur in the same areas across the 2 heatmaps. The longitude range of -73.96 to -73.94, and the latitude range 40.78 to 40.80. The areas that had darker hues in one heatmap, but not the other. Meaning in those areas a certain crime was committed increasingly more than the other crime.

Number 13

- 0-5 and 6-11:
  - During the 0-5 phase we see a high occurrence of crime being committed:
    - the longitude range of -74.00 to -73.96 and the latitude range of 40.70 to 40.80
    - the longitude range of -73.96 to -73.94 and the latitude range of 40.70 t0 40.72
    - the longitude range of -73.95 to -73.91 and the latitude range of 40.71 t0 40.77

We see an upward trend of crime volume being committed at higher latitude around the high crime frequency area of phase 0-5. We can also see that the crimes being committed is a lot more spread out in the later phase of 6-11

- 6-11 and 12-17
  - During the 6-11 phase, we see a trend of crimes being committed at low longitude locations all around the latitude areas. As we progress into the later phase 12-17, we see a gradual shift of crimes/harassments being committed a little bit higher longitude. We can also see less crime being committed at high longitude territory.