Homework 4 zl9901

1 a) It is not data mining task, because the classification is not based on linear regression or SVM. We can't say it's data mining according to gender.

b) It is data mining because it has made predictions.

c) It is data mining because it has trained data which is earthquake activities and use the trained data to make predictions.

d) It is not data mining because extract sample only can not be defined as data mining.

2 a) $\cos\theta = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = 1.0$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{0}{0}$$

In this case, we can't get the actual value of r

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = 2$$

b) $\cos\theta = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = 0$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = -1$$

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = 2$$

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = 0$$

c) $$\cos\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = 0$$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = 0$$

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} = 2$$

d) $$\cos\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = \frac{3}{4}$$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{1}{4}$$

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{3}{5}$$

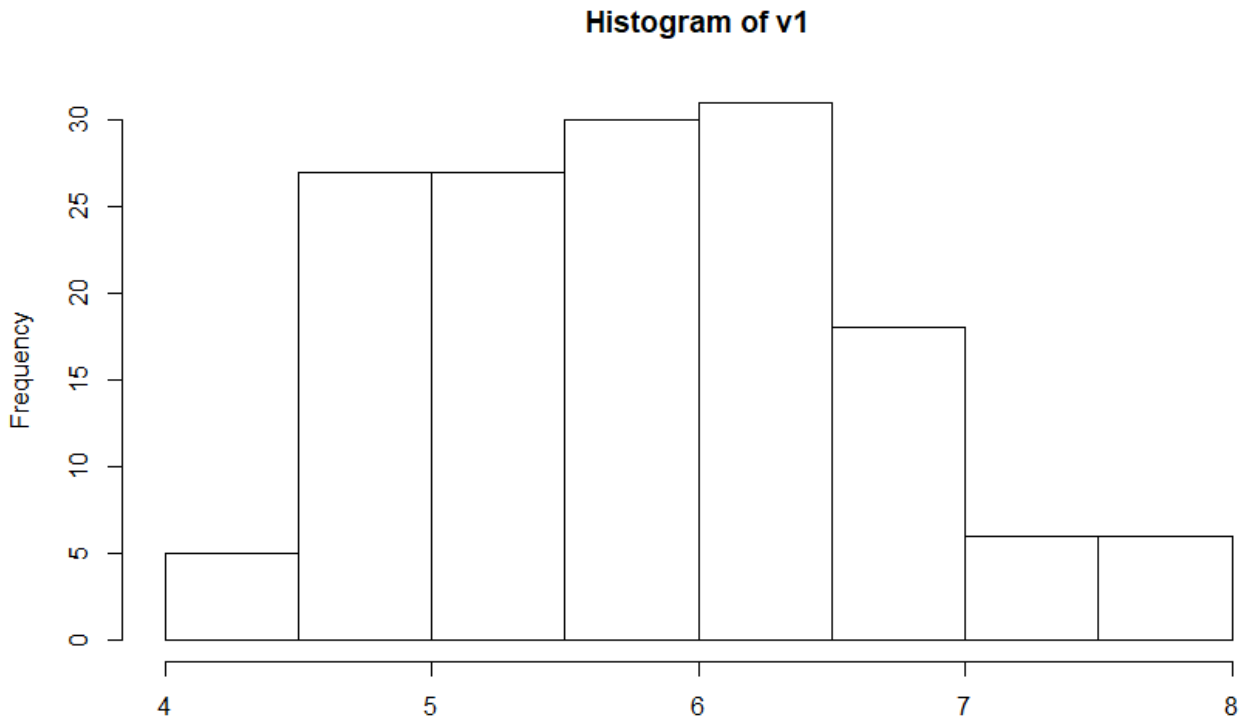
e) $$\cos\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = 0$$

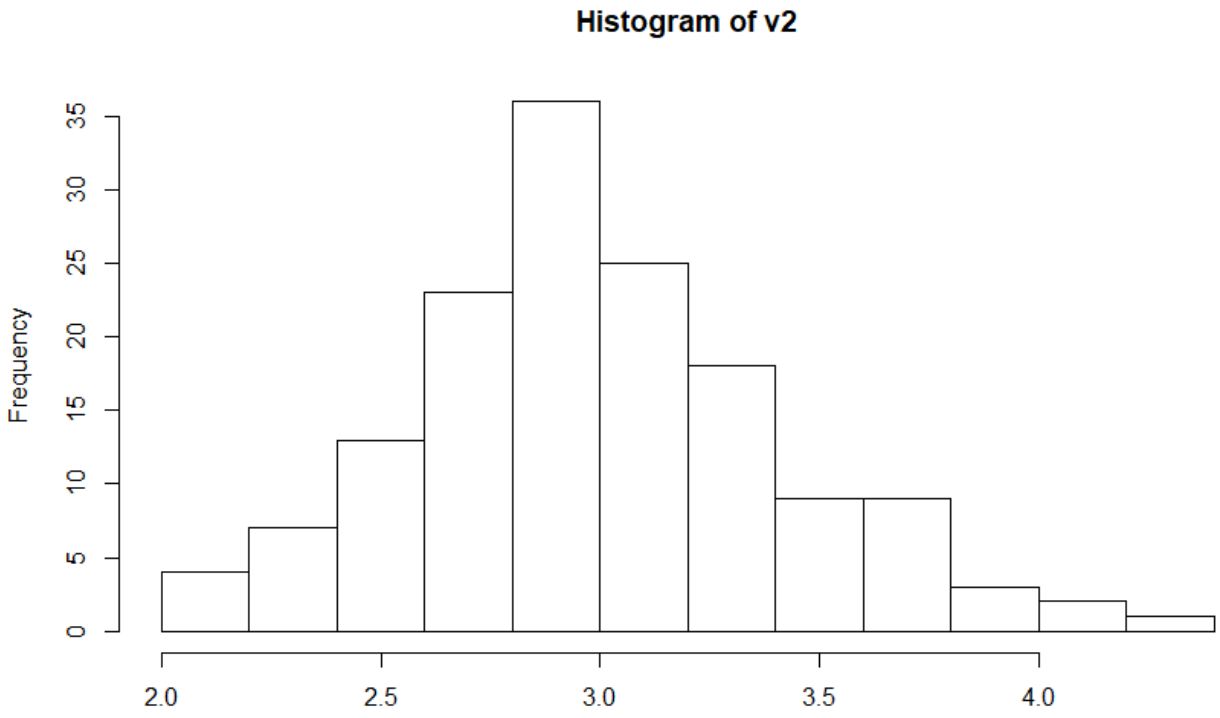
$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = 0$$

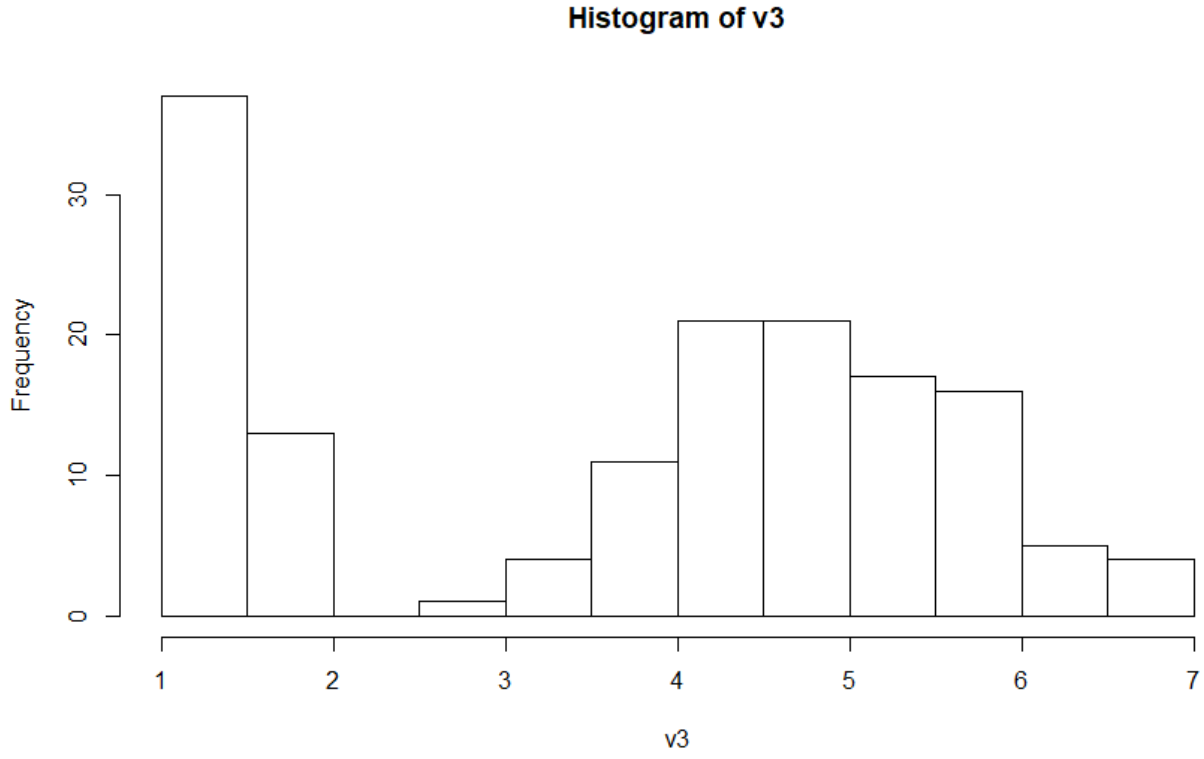
3

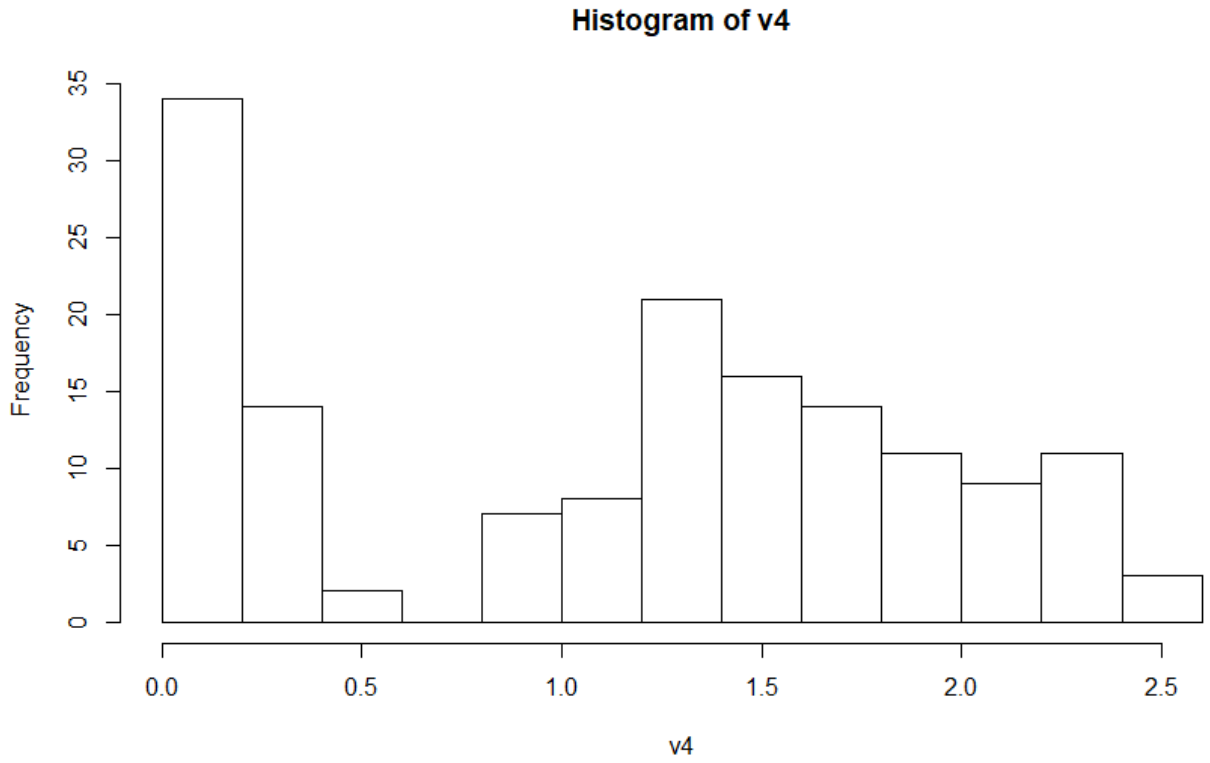
values	
cylinders	Factor w/ 5 levels "3","4","5","6",..: 5 5 5 5 5 5 5 5 5 5 ...
sepalLength	num [1:150] 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
sepalmean	5.84333333333333
sepalmedian	5.8
sepalpercentile	Named num [1:5] 4.3 5.1 5.8 6.4 7.9
sepalrange	num [1:2] 4.3 7.9
sepalvariance	0.685693512304251

```
> summary(iris)
       V1              V2              V3              V4              V5
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   Length:150
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   Class :character
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   Mode  :character
 Mean   :5.843   Mean   :3.054   Mean   :3.759   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

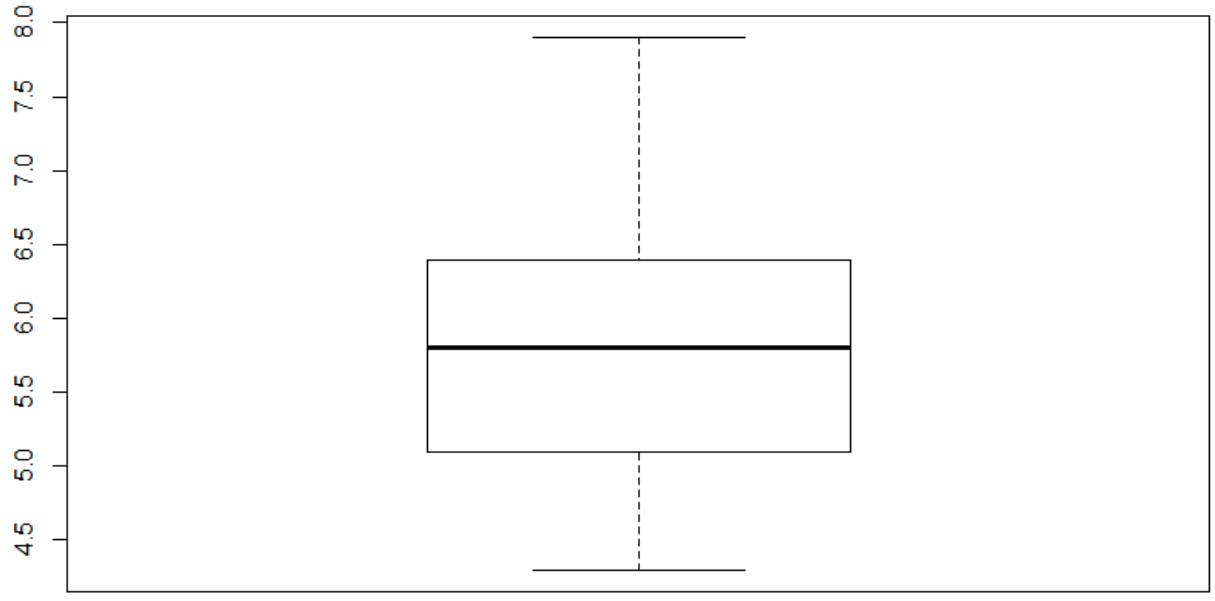



Histogram of v2
Frequency
0
5
10
15
20
25
30
35
2.0
2.5
3.0
3.5
4.0

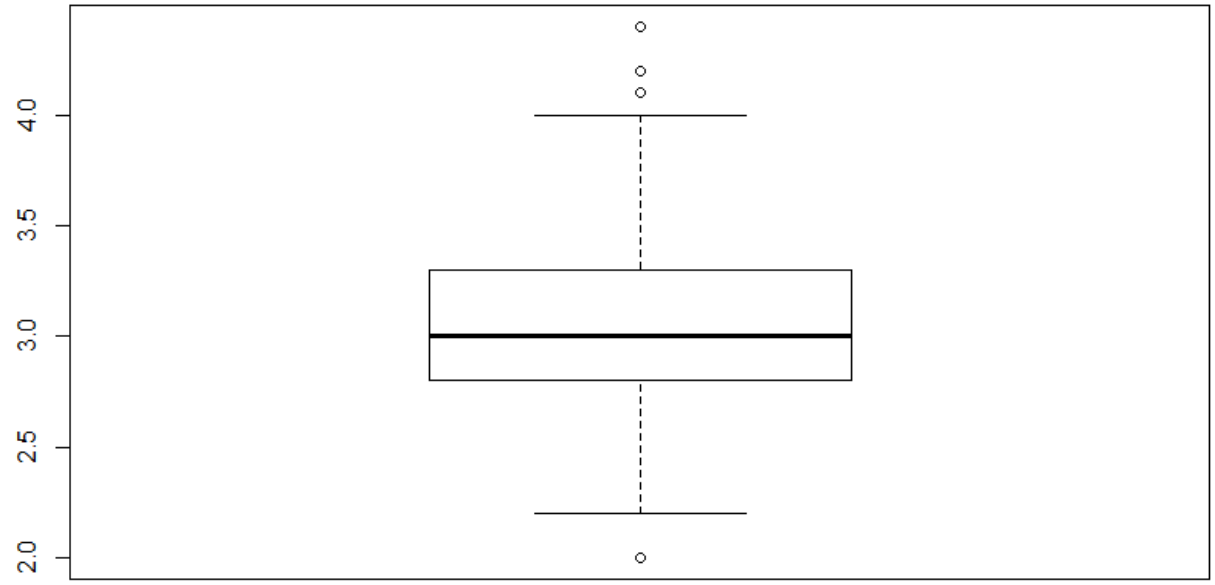

Histogram of v3
Frequency
0
10
20
30
1
2
3
4
5
6
7
v3



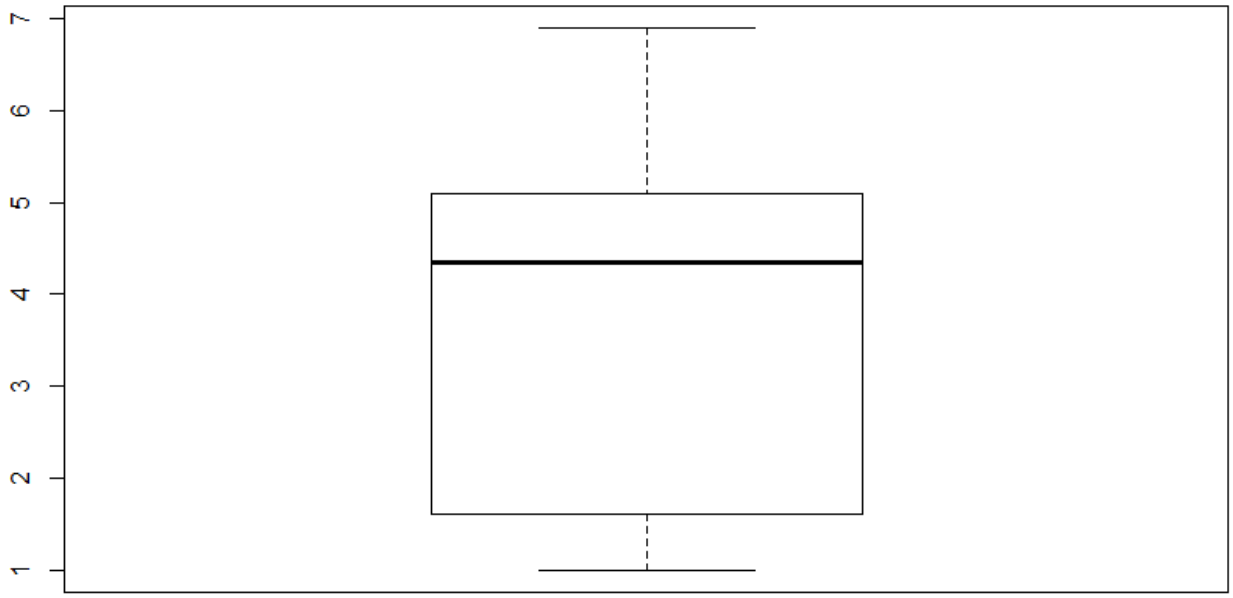
Box Plot of v1



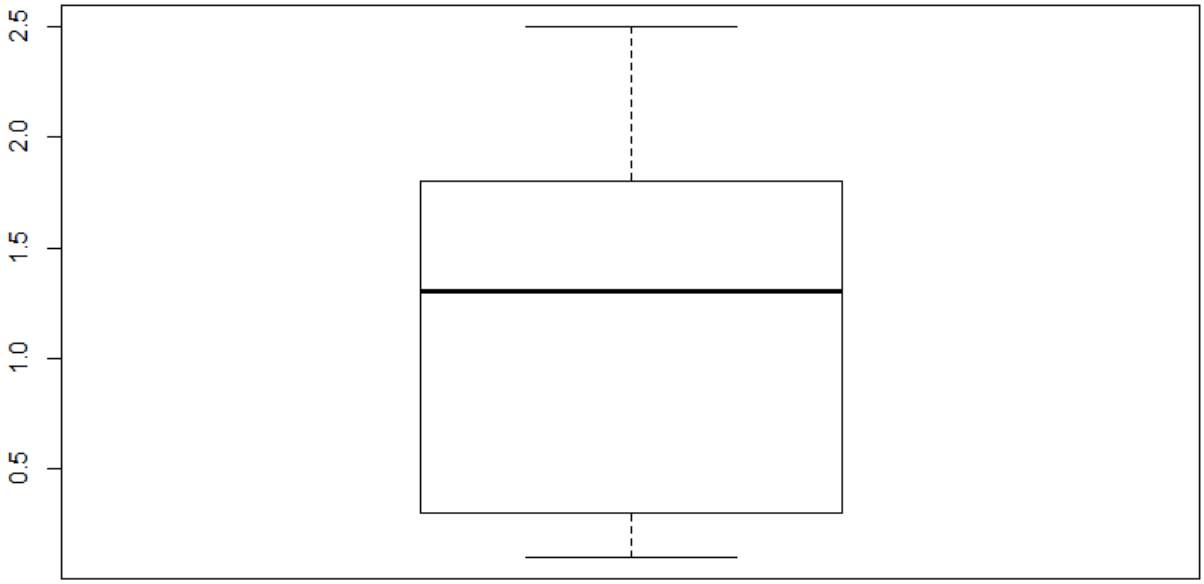
Box Plot of v2



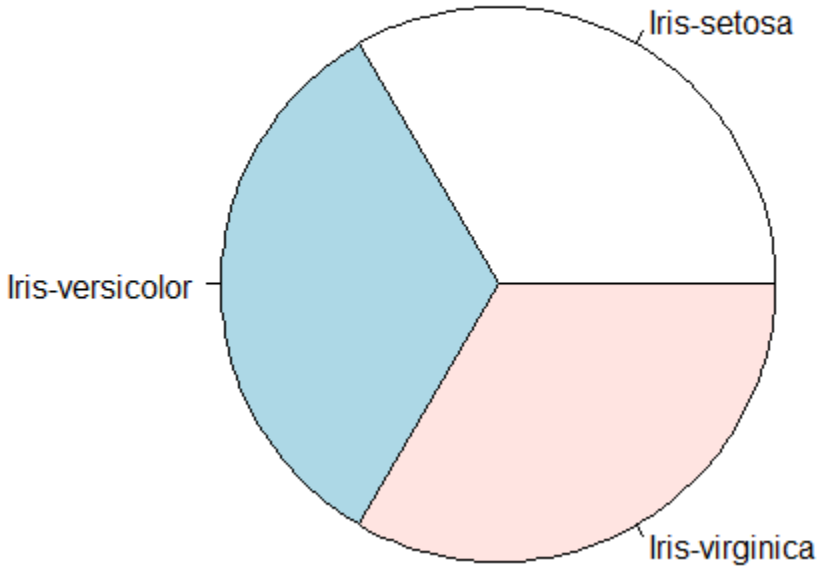
Box Plot of v3

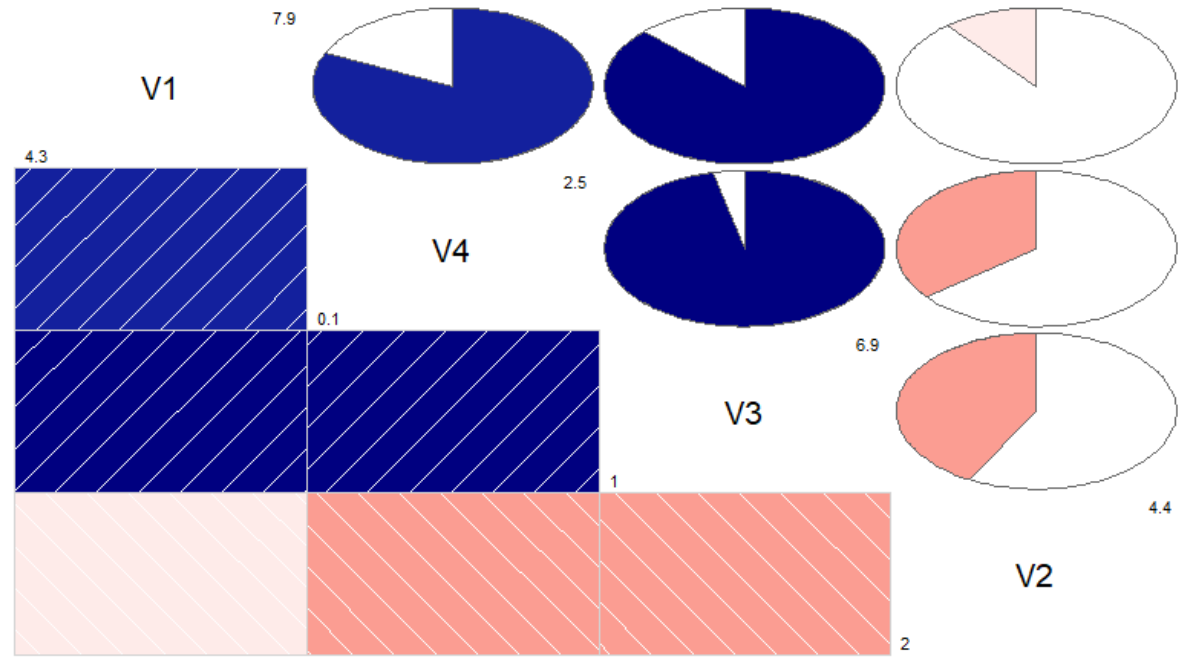


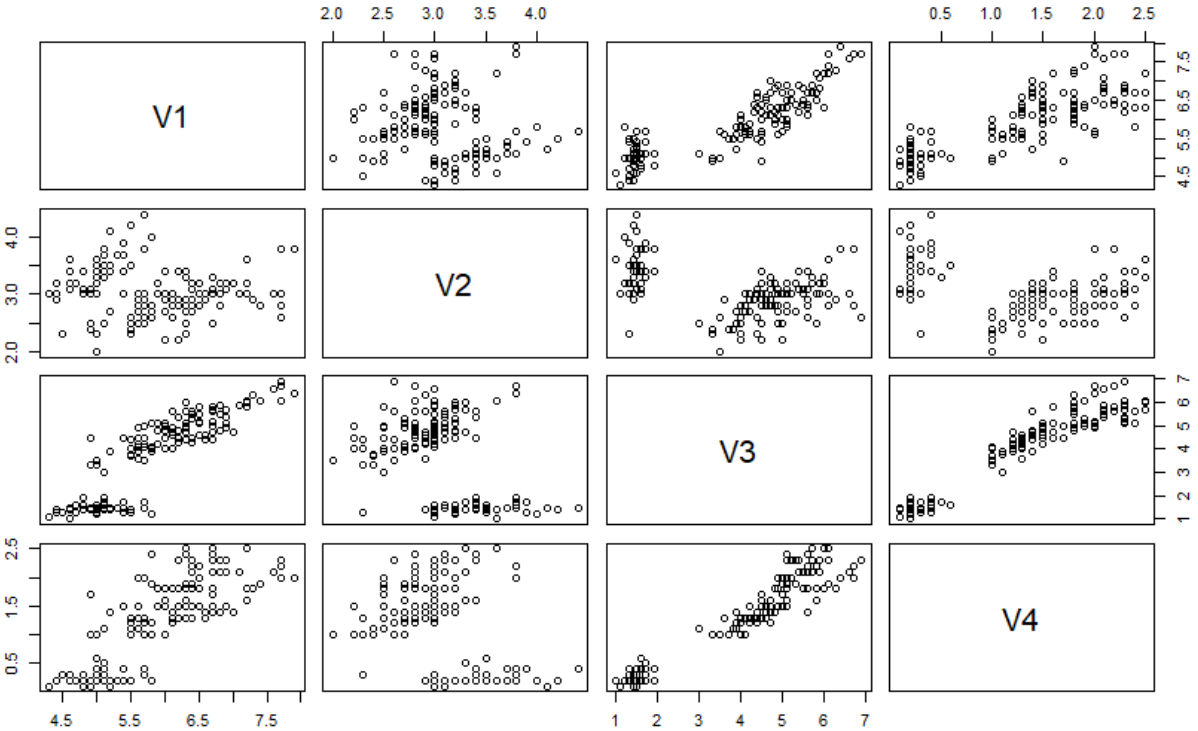
Box Plot of v4



Pie Chart of the Distribution of the types of Iris flowers




correlogram of mydata intercorrelations
V1
V4
V3
V2
7.9
4.3
2.5
0.1
6.9
1
4.4
2


V1
V2
V3
V4
2.0 2.5 3.0 3.5 4.0
0.5 1.0 1.5 2.0 2.5
4.5 5.5 6.5 7.5
2.0 3.0 4.0
1 2 3 4 5 6 7
0.5 1.5 2.5
4.5 5.5 6.5 7.5
1 2 3 4 5 6 7

```r
require(data.table)

iris <-as.data.frame(fread("iris.data", quote = ""))
sepalLength <-iris[,1]

sepalmean <-mean(sepalLength)
print(mean)
sepalmedian <-median(sepalLength)
print(sepalmedian)
sepalrange <-range(sepalLength)
print(sepalrange)
sepalvariance <-var(sepalLength)
print(sepalvariance)
sepalpercentile <-quantile(sepalLength)
print(sepalpercentile)

summary(iris)

v1 <-iris[,1]
hist(v1)
v2 <-iris[,2]
hist(v2)
v3 <-iris[,3]
hist(v3)
v4 <-iris[,4]
hist(v4)
v5 <-iris[,5]

boxplot(v1)
boxplot(v2)
boxplot(v3)
boxplot(v4)

test <-table(v5)
num <-as.numeric(test)
pie(num,names(test))

library(corrgram)
corrgram(iris[,1:4],order=TRUE,lower.panel=panel.shade,
         upper.panel=panel.pie,text.panel=panel.txt,
         main='correlogram of mydata intercorrelations',
         diag.panel=panel.minmax
           )

pairs(~V1+V2+V3+V4,iris)
```