lista 2.

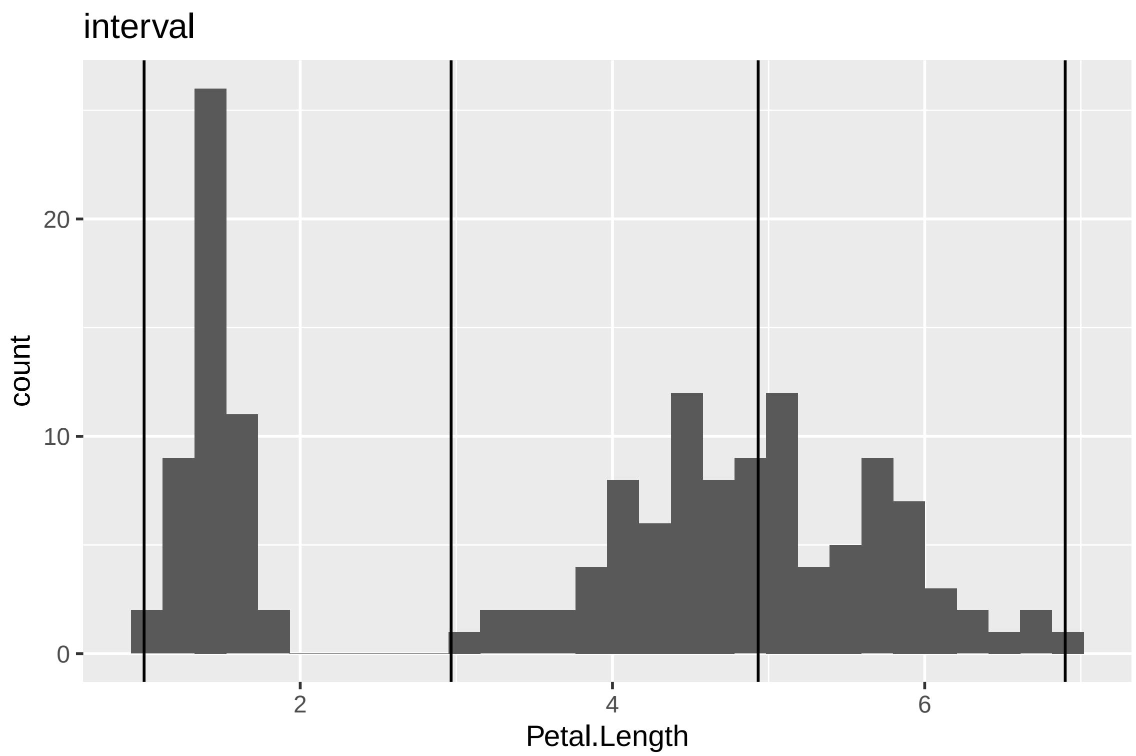
Mikołaj Langner, Marcin Kostrzewa

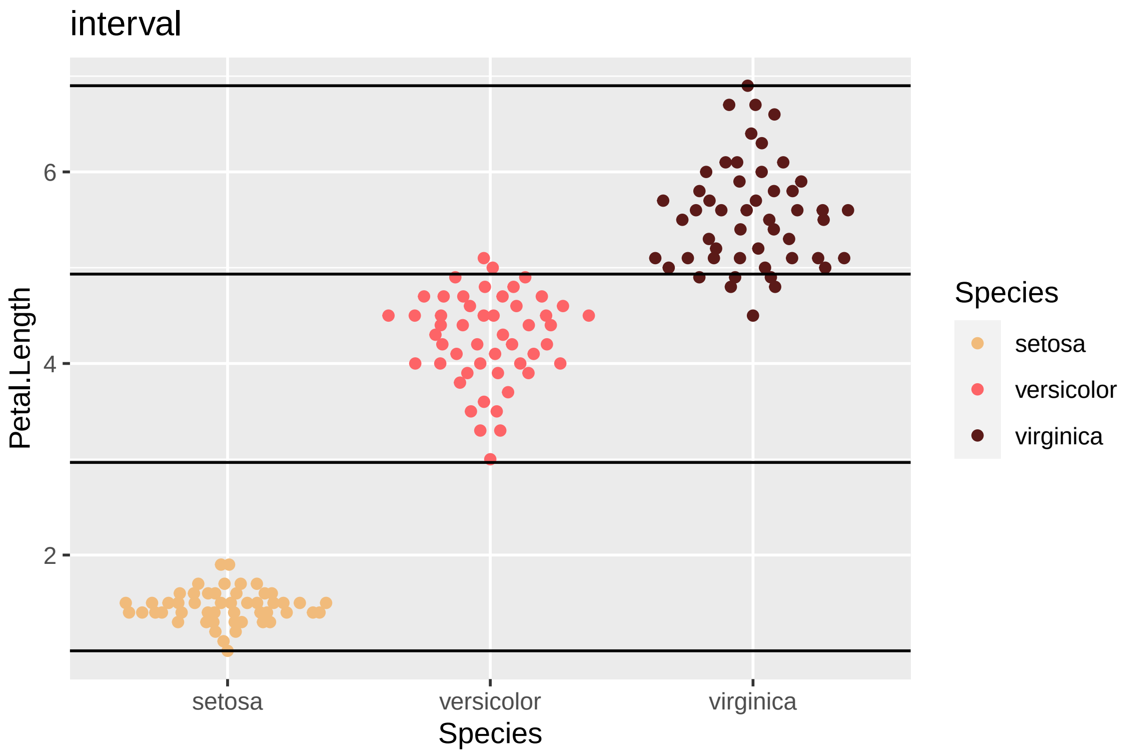
25.04.2021

ZADANIE 1.

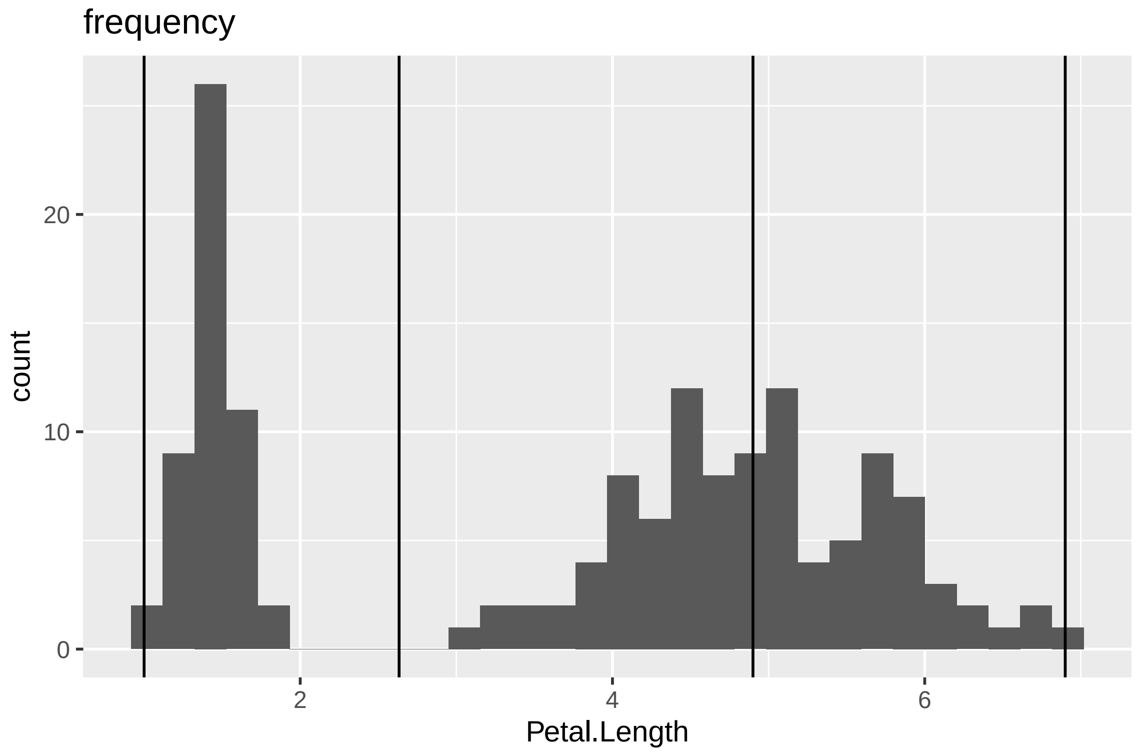
```
data(iris)
```

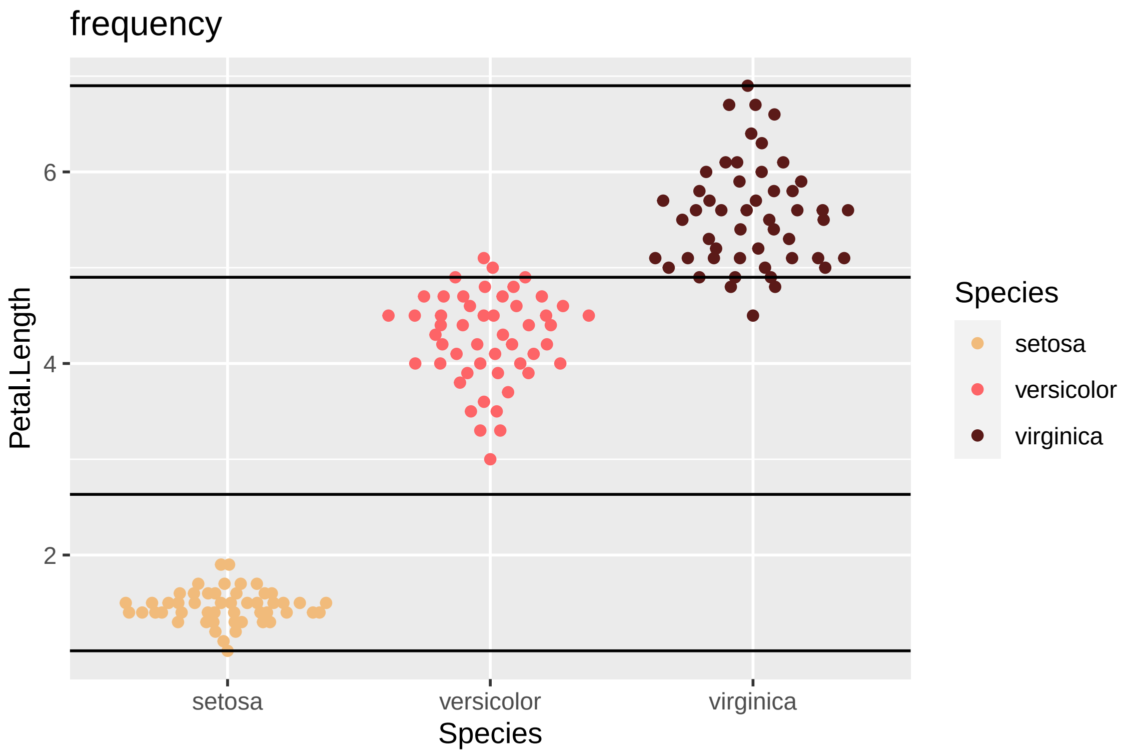
```
intervals <- c(min(iris$Petal.Length), 2, 5, max(iris$Petal.Length))
for (method in c("interval", "frequency", "cluster", "fixed")) {
  petal.length.discretized <- if (method != "fixed")
    discretize(iris$Petal.Length, method=method) else
    discretize(iris$Petal.Length, method=method, breaks=intervals)
  print(ggplot(iris, aes(Petal.Length)) +
          geom_histogram() +
          geom_vline(xintercept=attributes(petal.length.discretized)$"discretized:breaks") +
          ggtitle(method))
  print(ggplot(iris, aes(Species, Petal.Length)) +
    geom_quasirandom(aes(col=Species)) +
    scale_color_manual(values=wes_palette("GrandBudapest1", 3)) +
    geom_hline(yintercept=attributes(petal.length.discretized)$"discretized:breaks") +
    ggtitle(method))
  discretized.table <- table(petal.length.discretized, iris$Species)
  matchClasses(discretized.table)
}
```



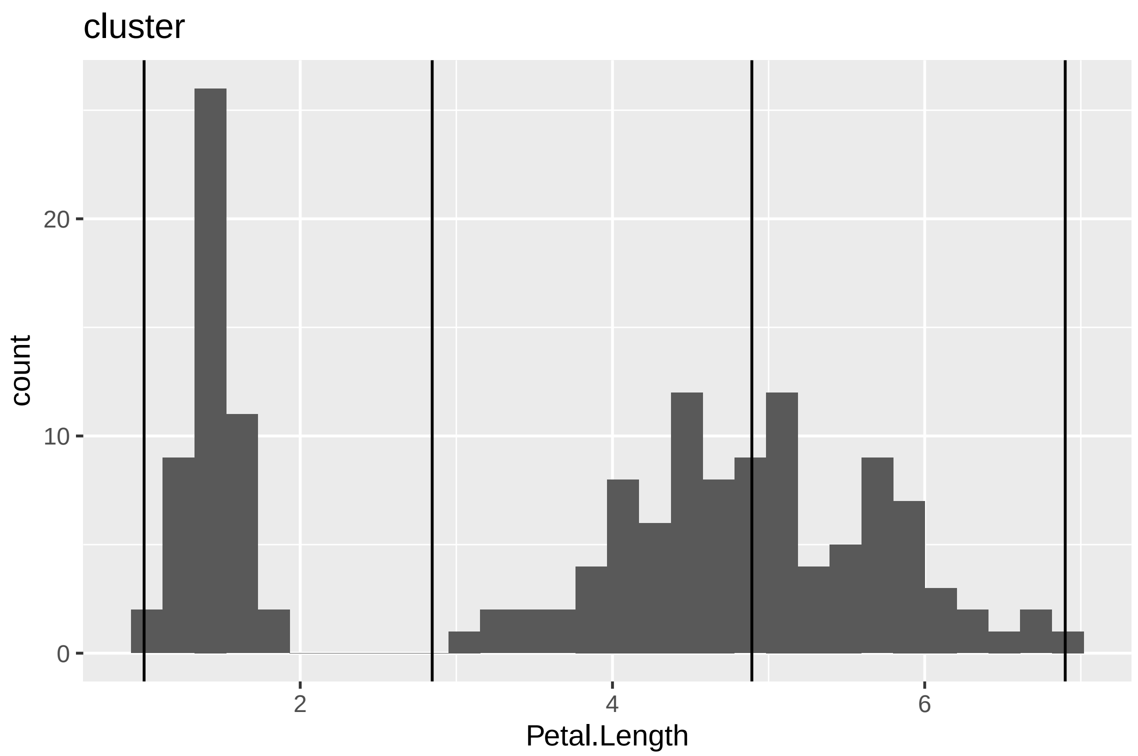


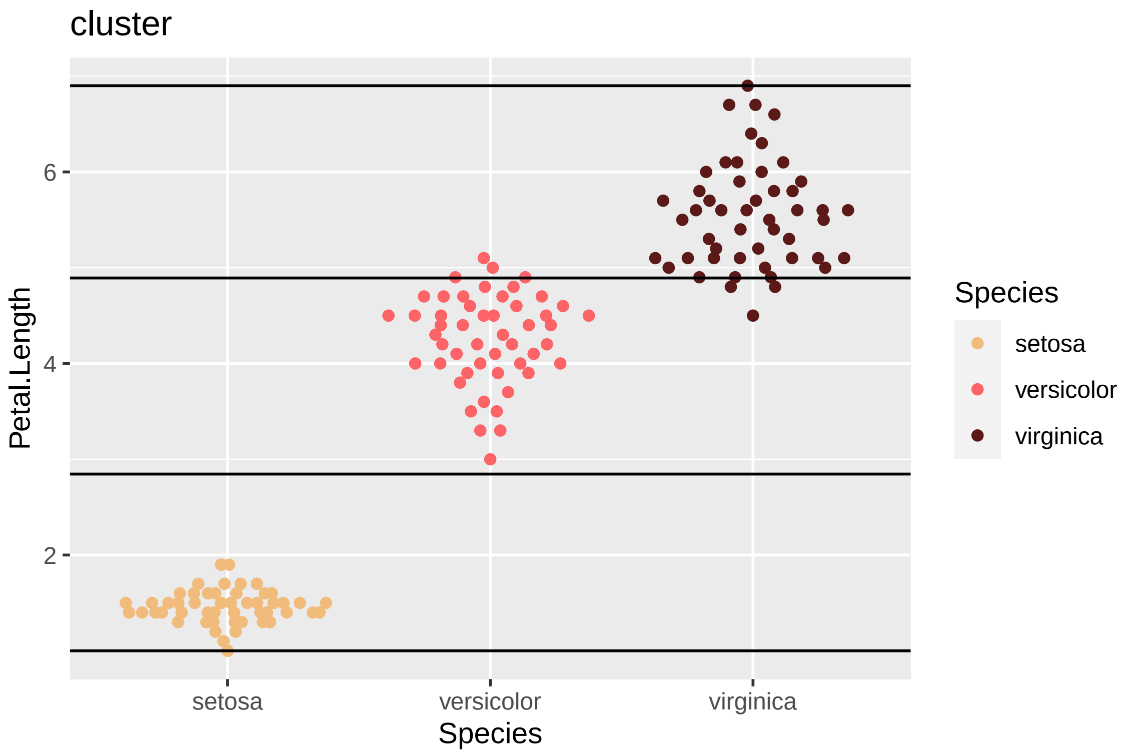
```
## Cases in matched pairs: 94.67 %
```



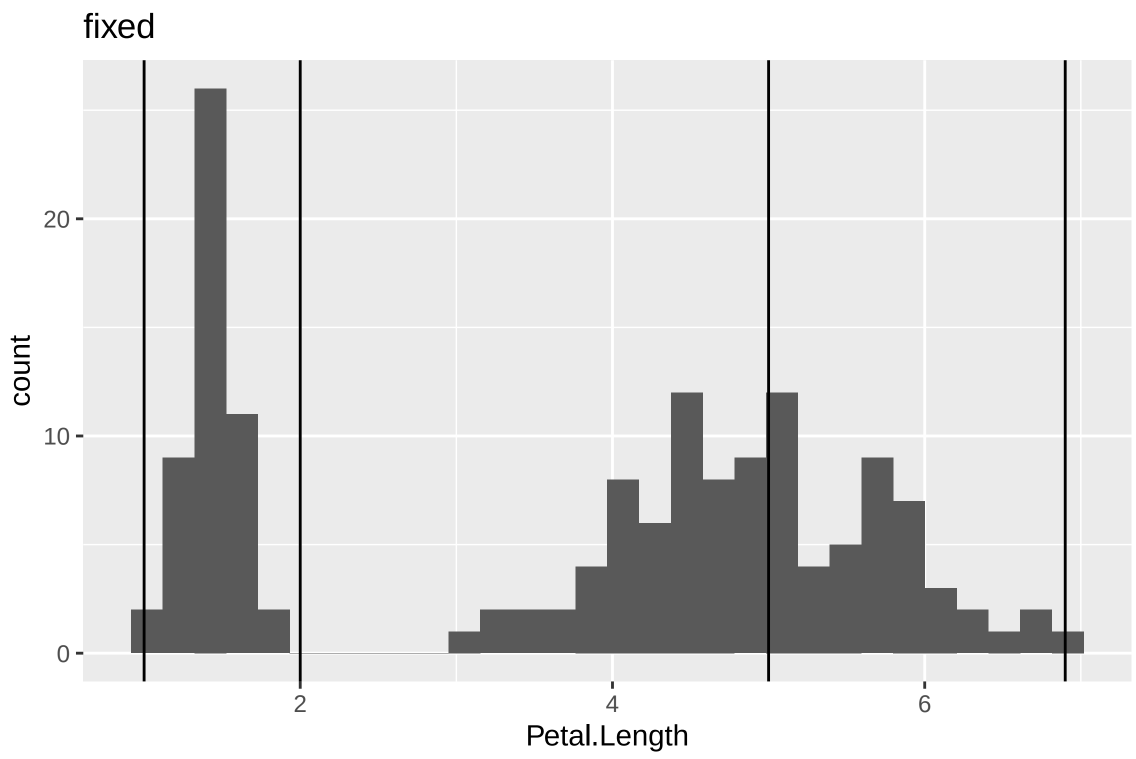


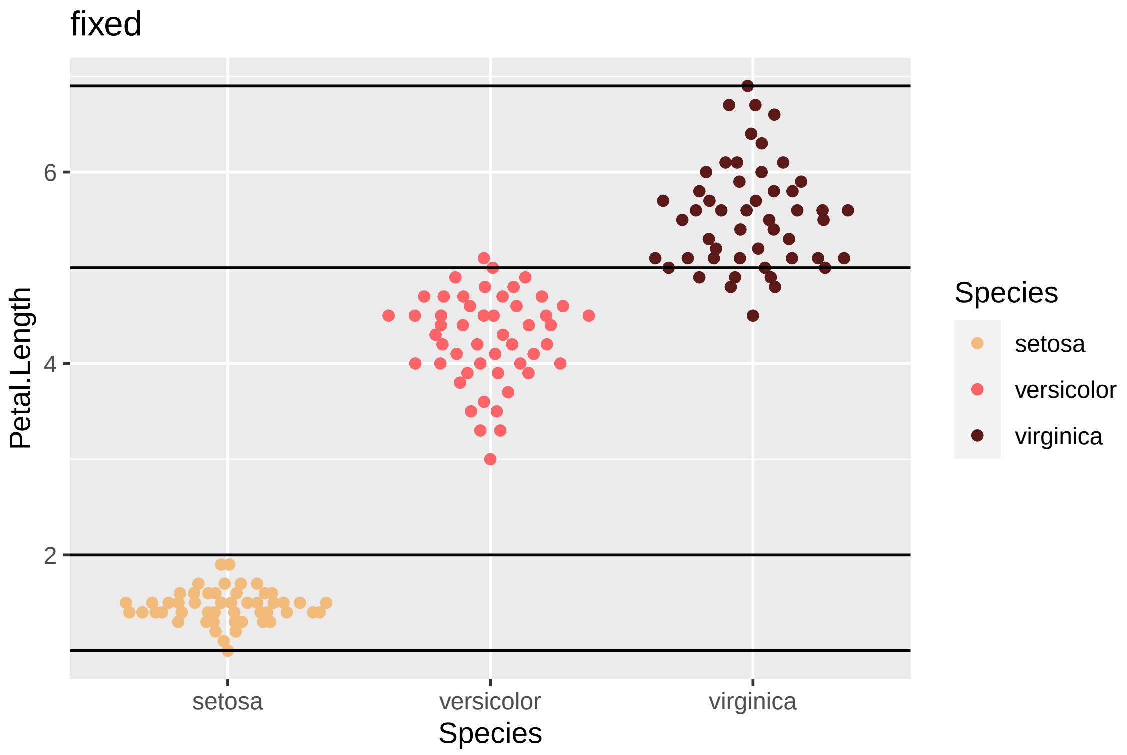
```
## Cases in matched pairs: 95.33 %
```





```
## Cases in matched pairs: 95.33 %
```

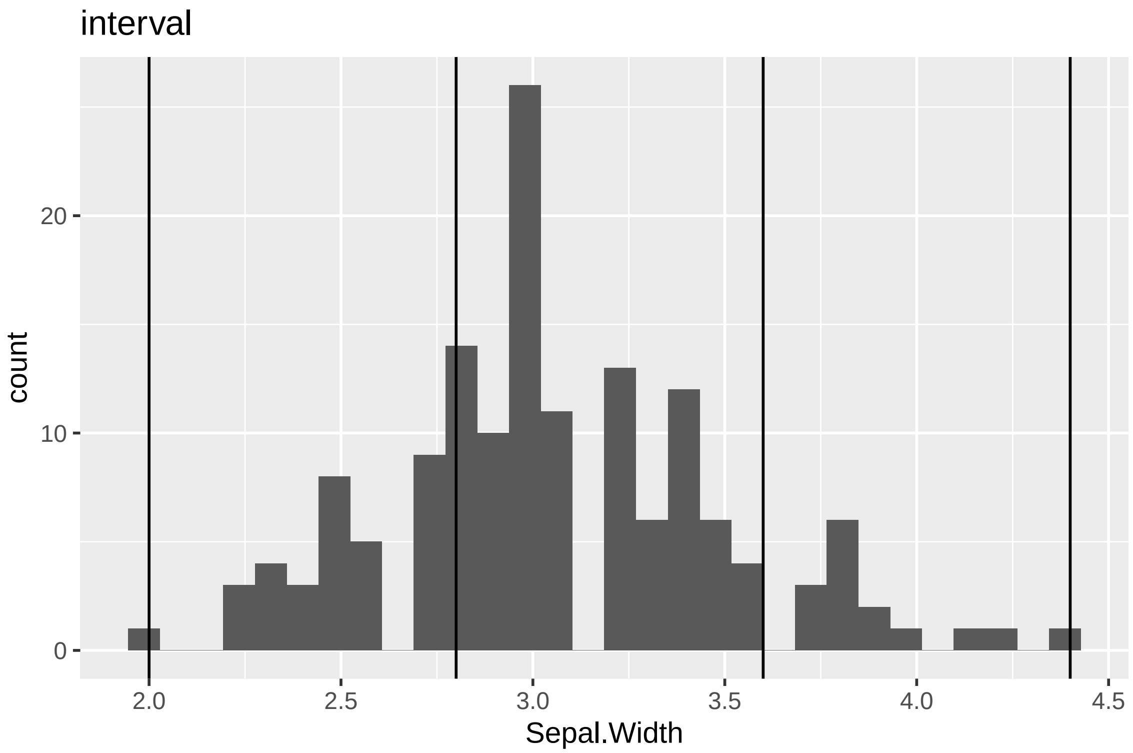


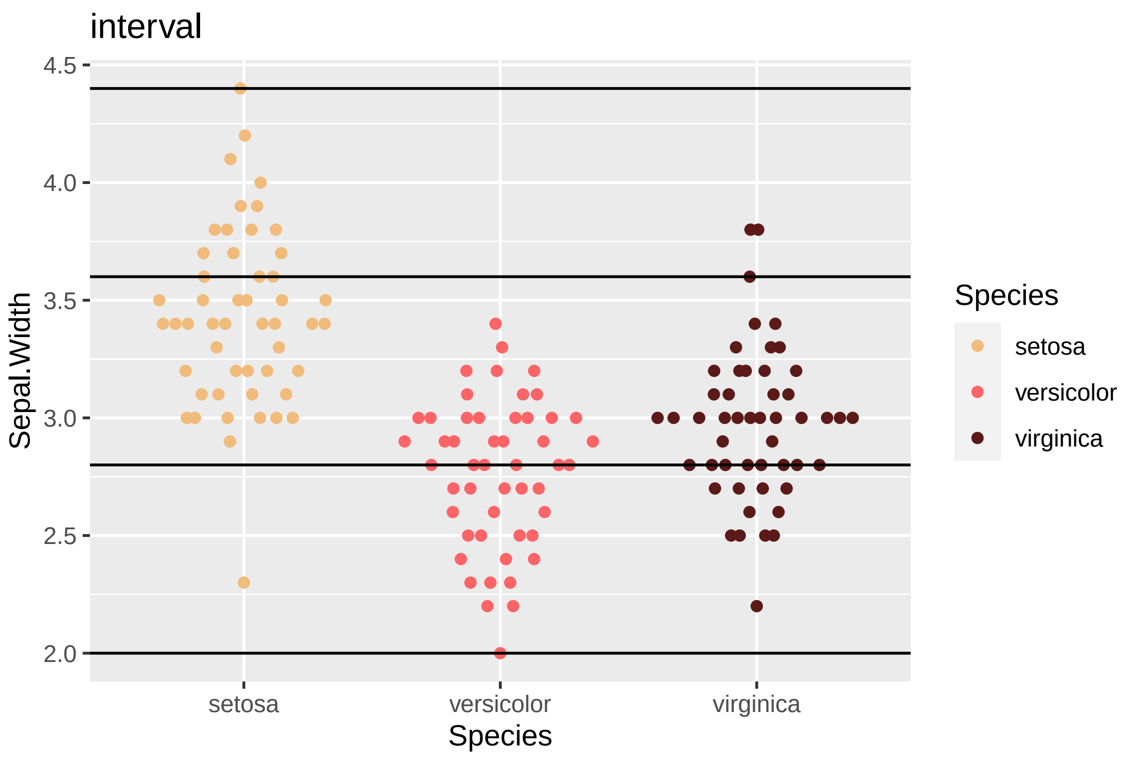


```
## Cases in matched pairs: 94.67 %
```

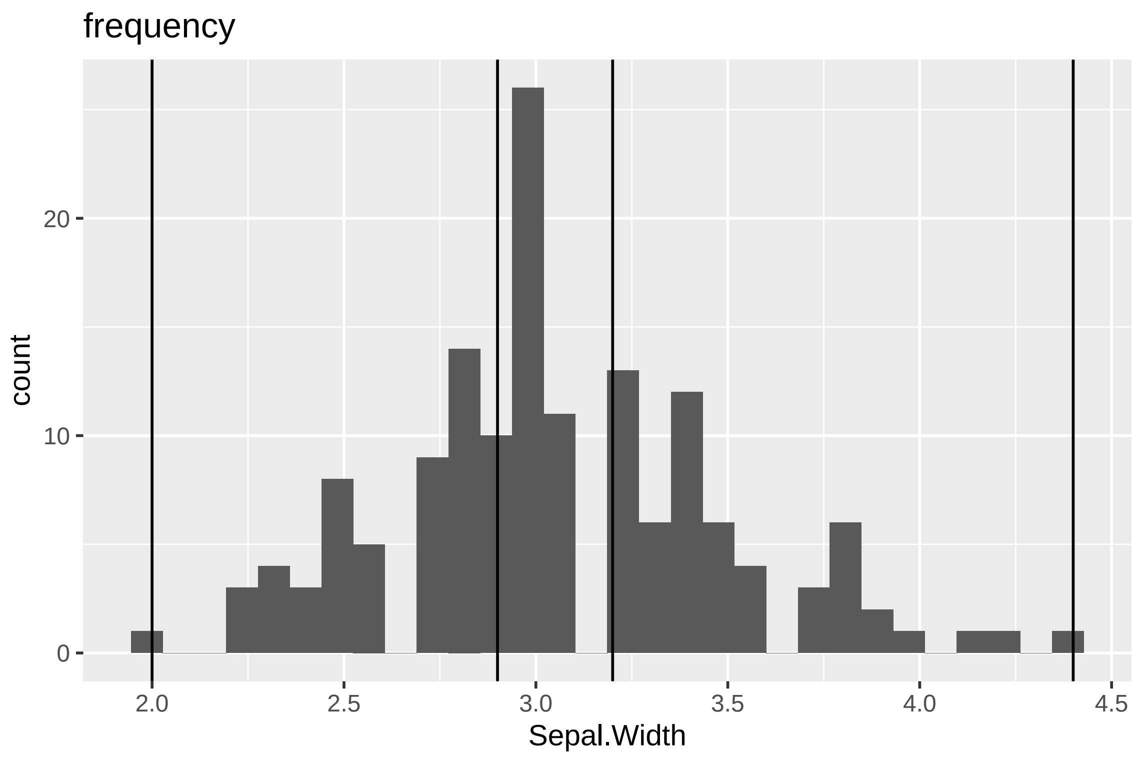
```
intervals <- c(min(iris$Sepal.Width), 2.5, 3, max(iris$Sepal.Width))
for (method in c("interval", "frequency", "cluster", "fixed")) {
  sepal.width.discretized <- if (method != "fixed")
```

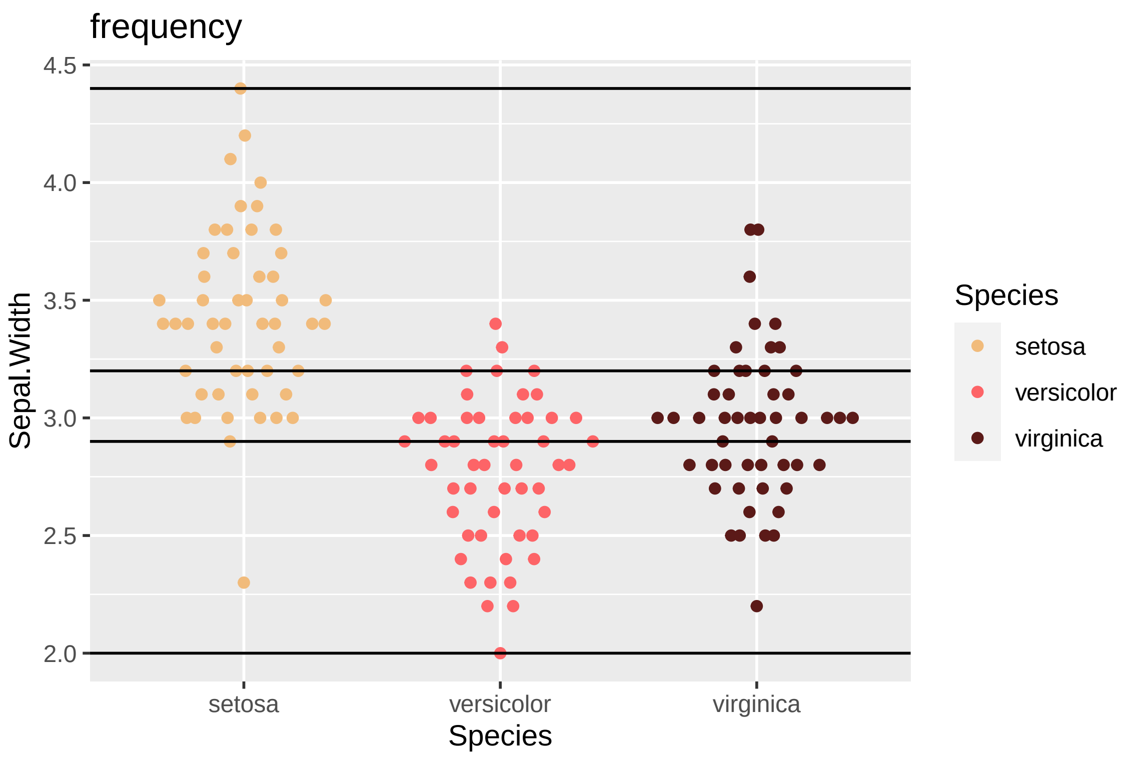
```
    discretize(iris$Sepal.Width, method=method) else
    discretize(iris$Sepal.Width, method=method, breaks=intervals)
  print(ggplot(iris, aes(Sepal.Width)) +
          geom_histogram() +
          geom_vline(xintercept=attributes(sepal.width.discretized)$"discretized:breaks") +
          ggtitle(method))
  print(ggplot(iris, aes(Species, Sepal.Width)) +
    geom_quasirandom(aes(col=Species)) +
    scale_color_manual(values=wes_palette("GrandBudapest1", 3)) +
    geom_hline(yintercept=attributes(sepal.width.discretized)$"discretized:breaks") +
    ggtitle(method))
  discretized.table <- table(sepal.width.discretized, iris$Species)
  matchClasses(discretized.table)
}
```



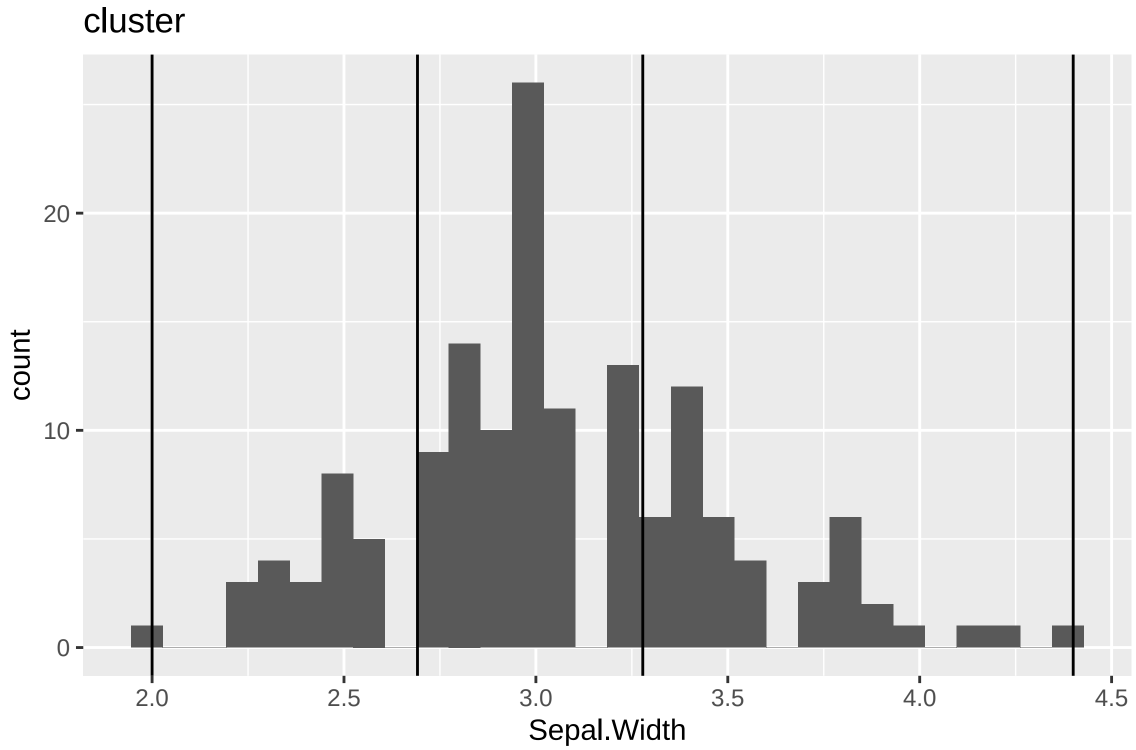


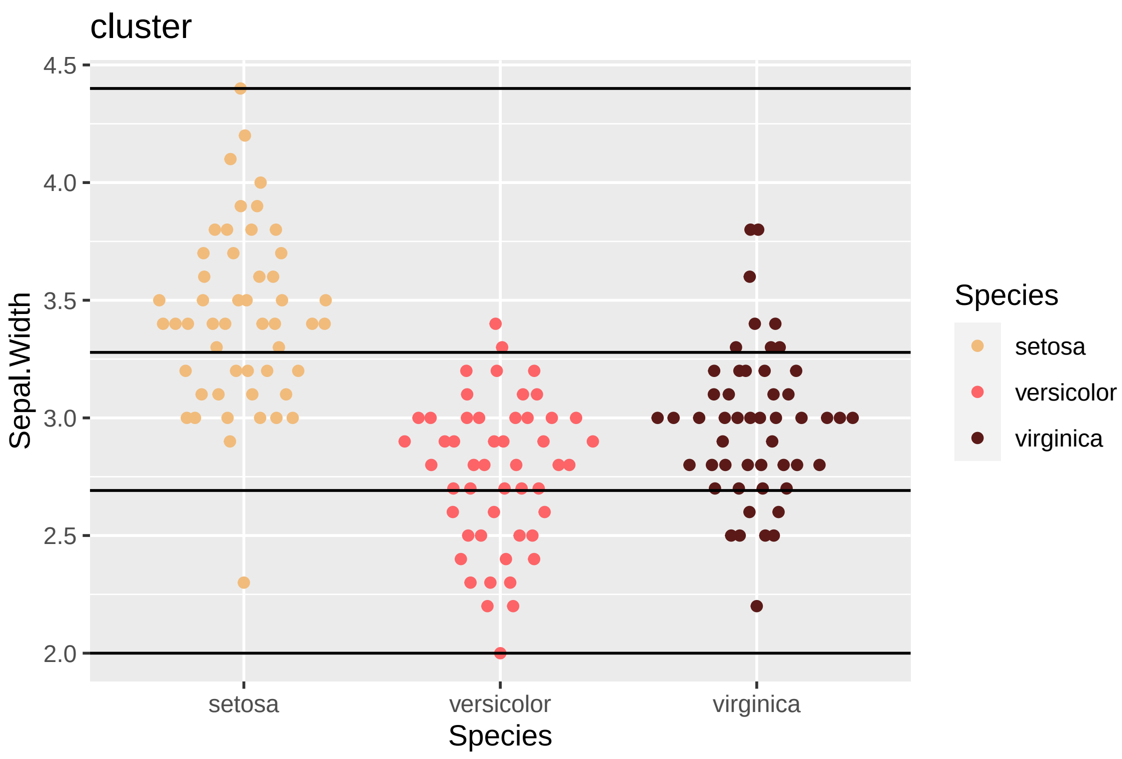
```
## Cases in matched pairs: 50.67 %
```



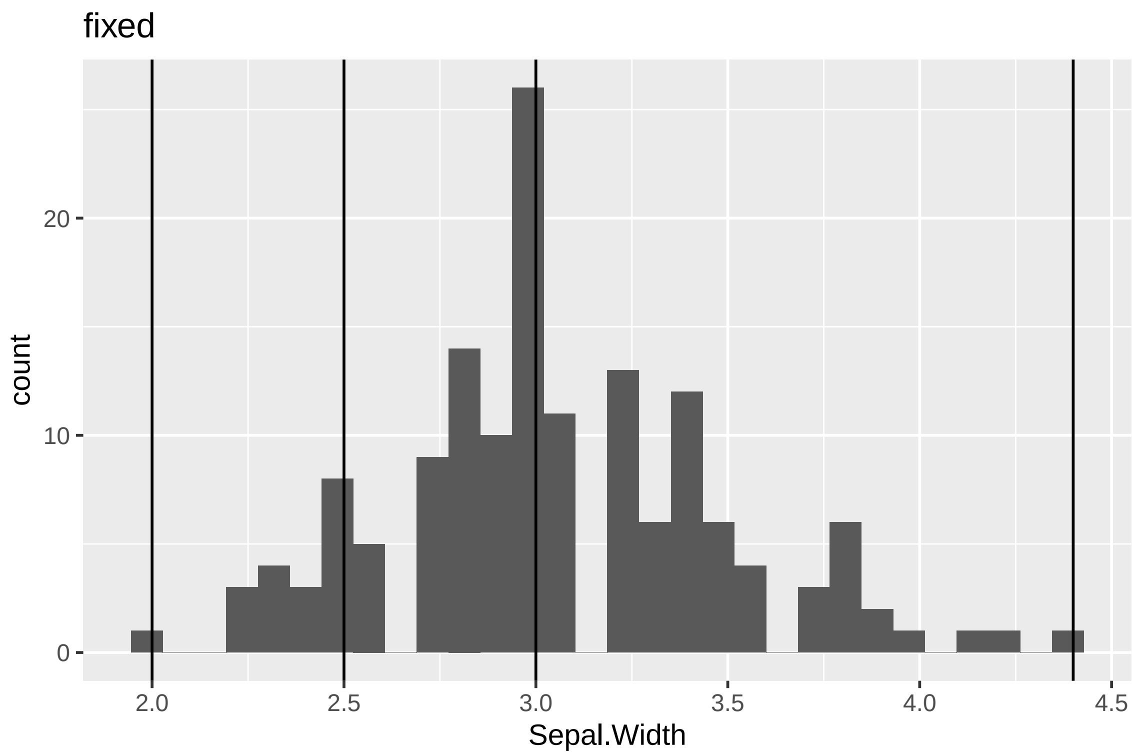


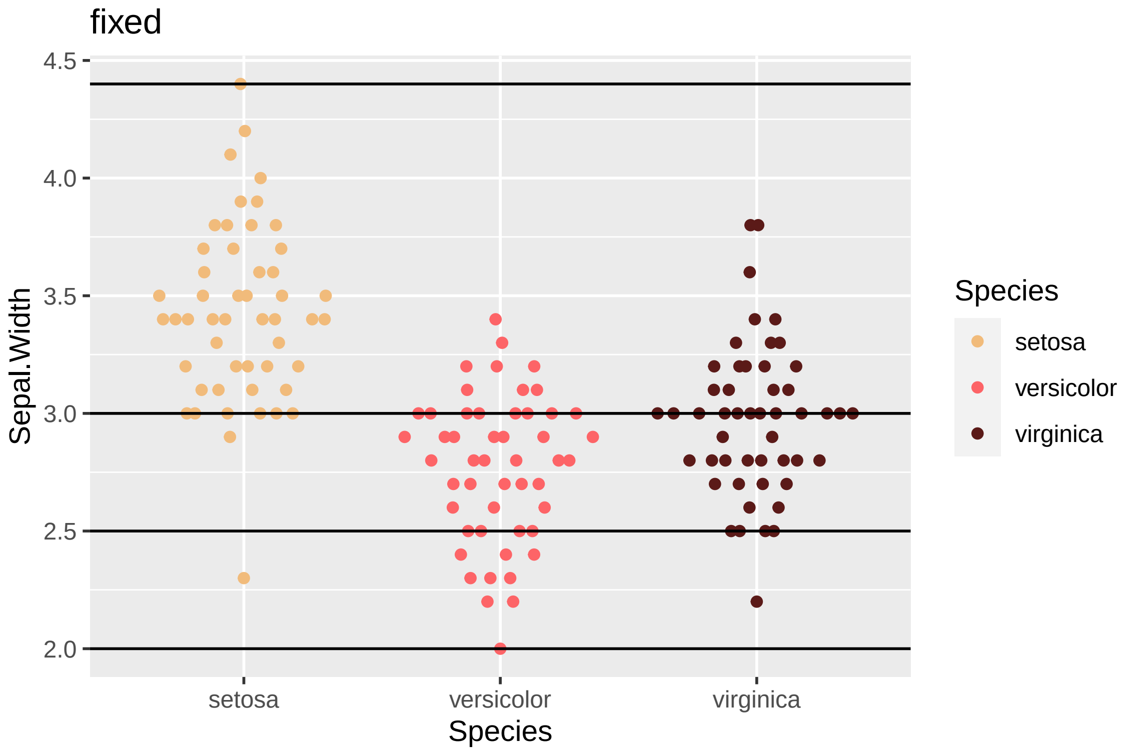
```
## Cases in matched pairs: 55.33 %
```





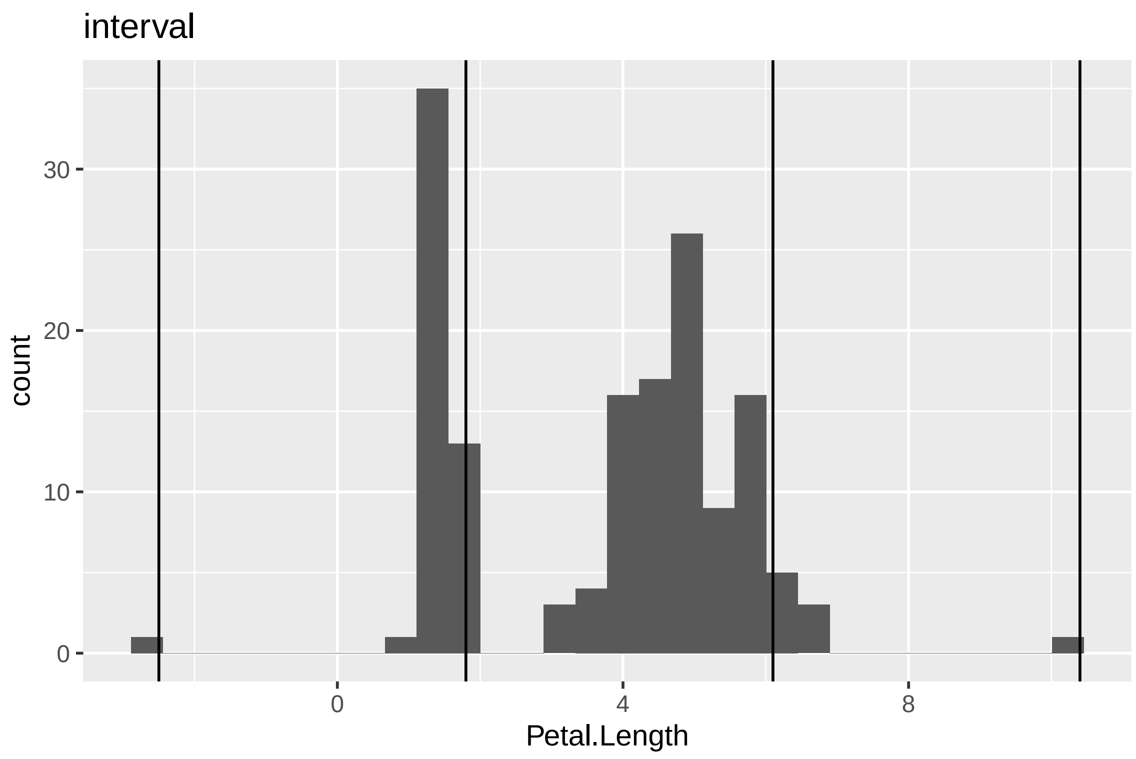
```
## Cases in matched pairs: 56 %
```

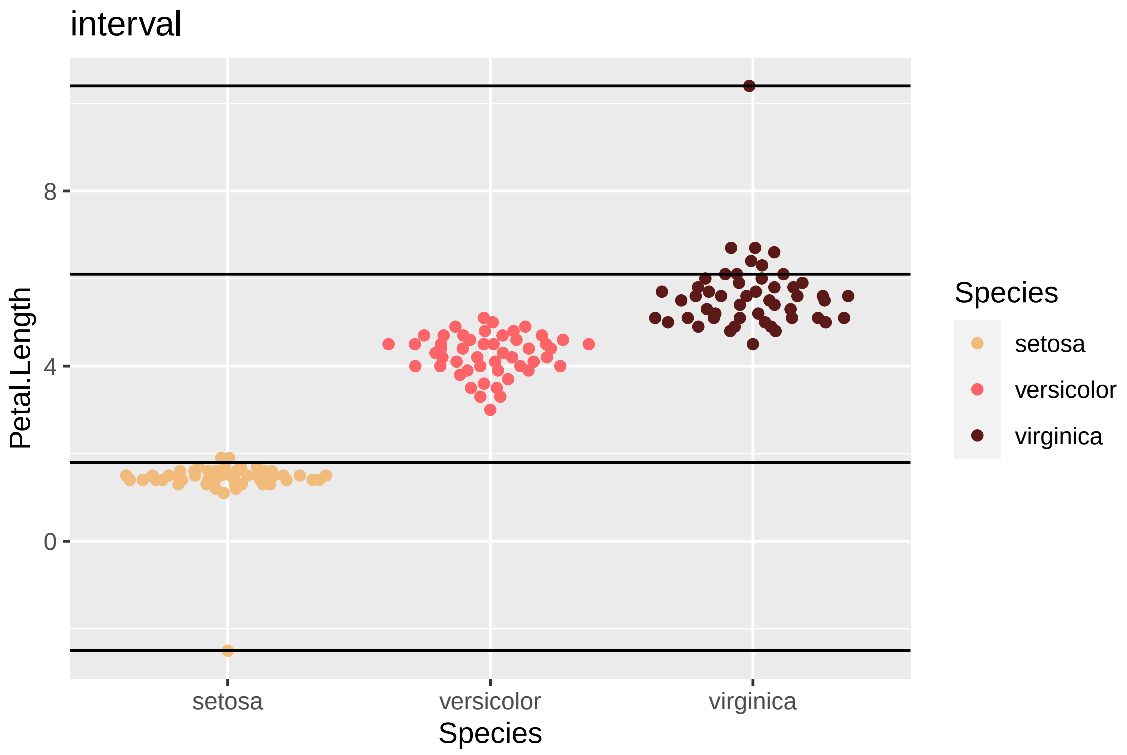




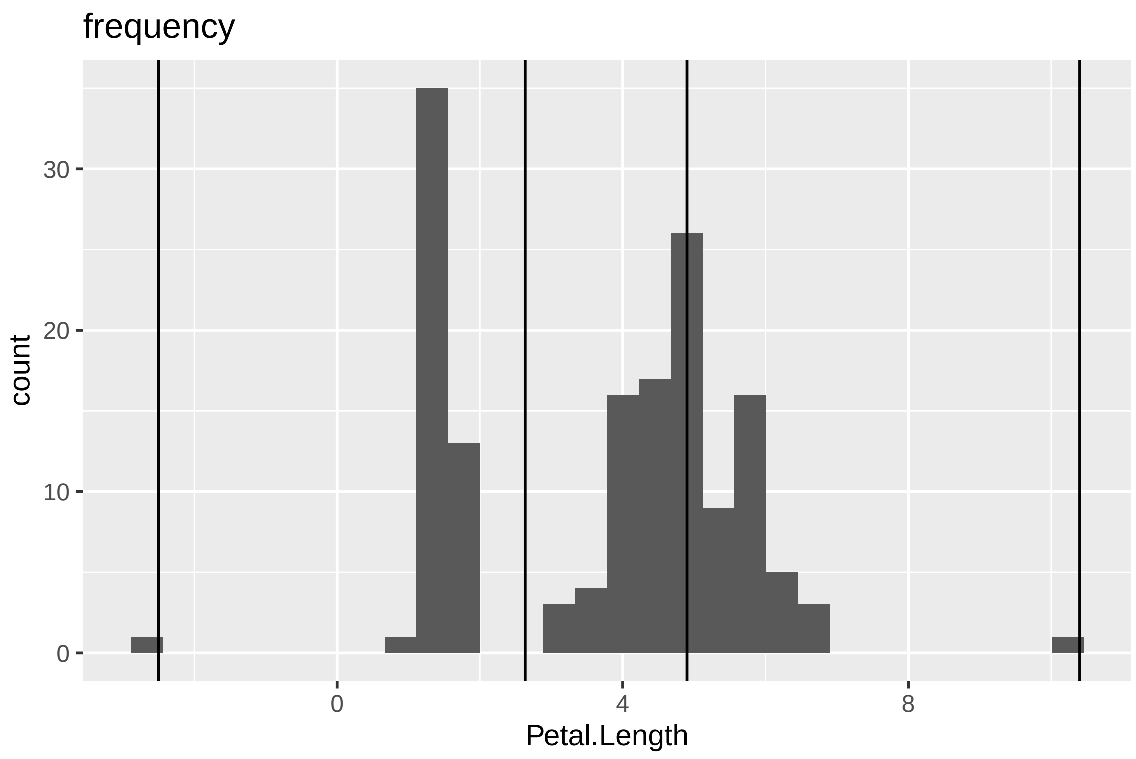
```
## Cases in matched pairs: 54.67 %
```

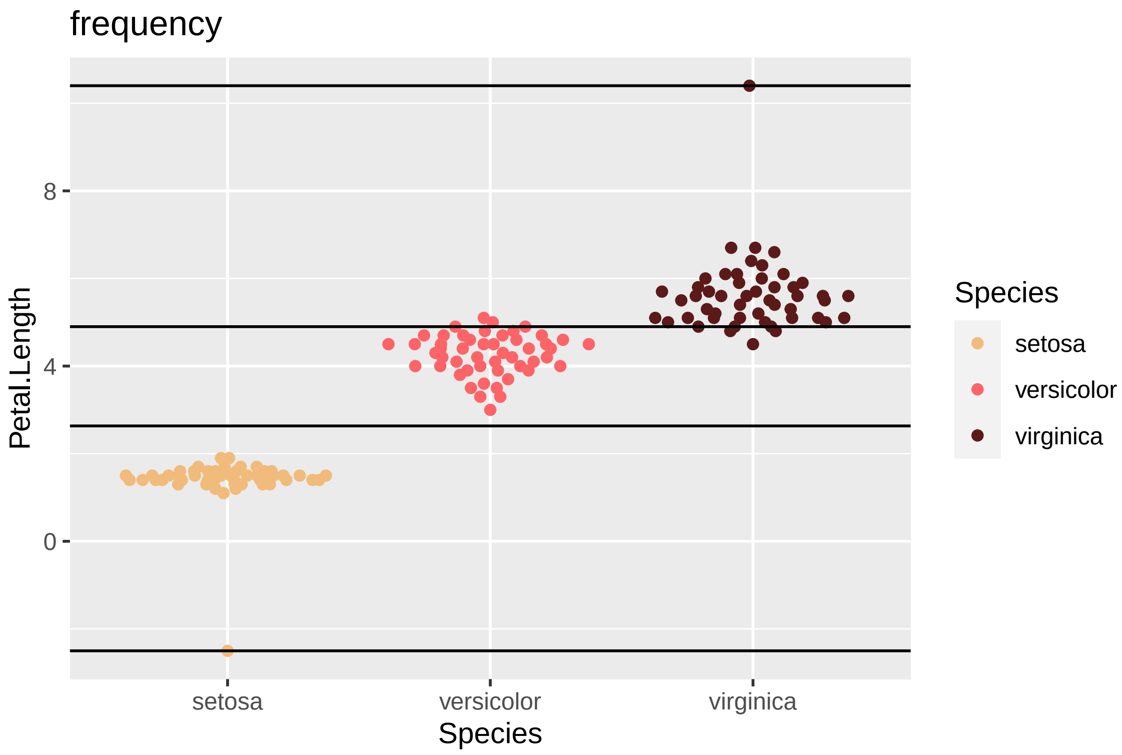
```
iris$Petal.Length[which.min(iris$Petal.Length)] <- min(iris$Petal.Length) - IQR(iris$Petal.Length)
iris$Petal.Length[which.max(iris$Petal.Length)] <- max(iris$Petal.Length) + IQR(iris$Petal.Length)
intervals <- c(min(iris$Petal.Length), 2, 5, max(iris$Petal.Length))
for (method in c("interval", "frequency", "cluster", "fixed")) {
  petal.length.discretized <- if (method != "fixed")
    discretize(iris$Petal.Length, method=method) else
    discretize(iris$Petal.Length, method=method, breaks=intervals)
  print(ggplot(iris, aes(Petal.Length)) +
          geom_histogram() +
          geom_vline(xintercept=attributes(petal.length.discretized)$"discretized:breaks") +
          ggtitle(method))
  print(ggplot(iris, aes(Species, Petal.Length)) +
    geom_quasirandom(aes(col=Species)) +
    scale_color_manual(values=wes_palette("GrandBudapest1", 3)) +
    geom_hline(yintercept=attributes(petal.length.discretized)$"discretized:breaks") +
    ggtitle(method))
  discretized.table <- table(petal.length.discretized, iris$Species)
  matchClasses(discretized.table)
}
```



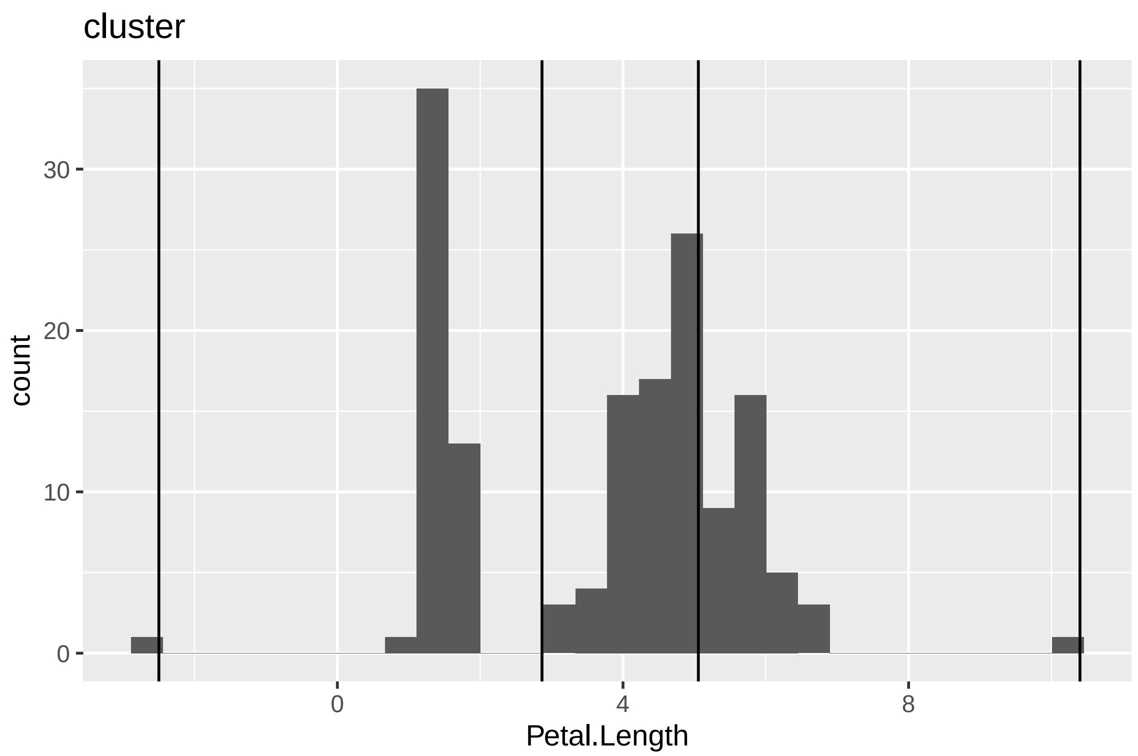


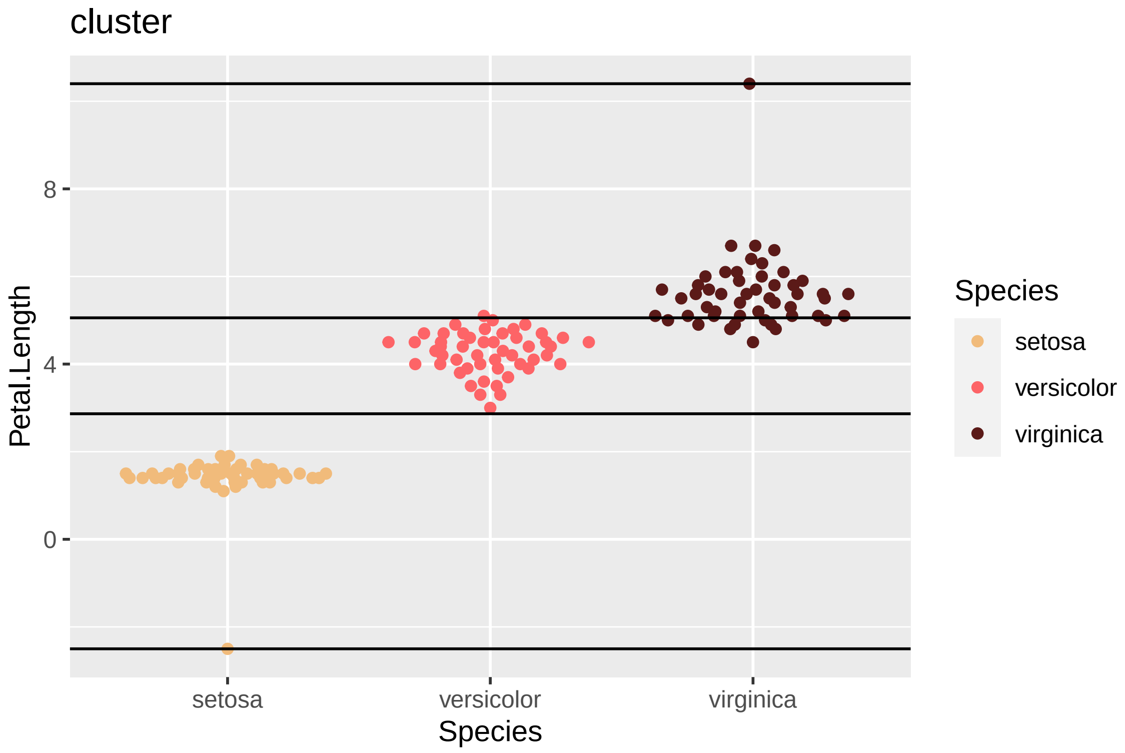
```
## Cases in matched pairs: 71.33 %
```



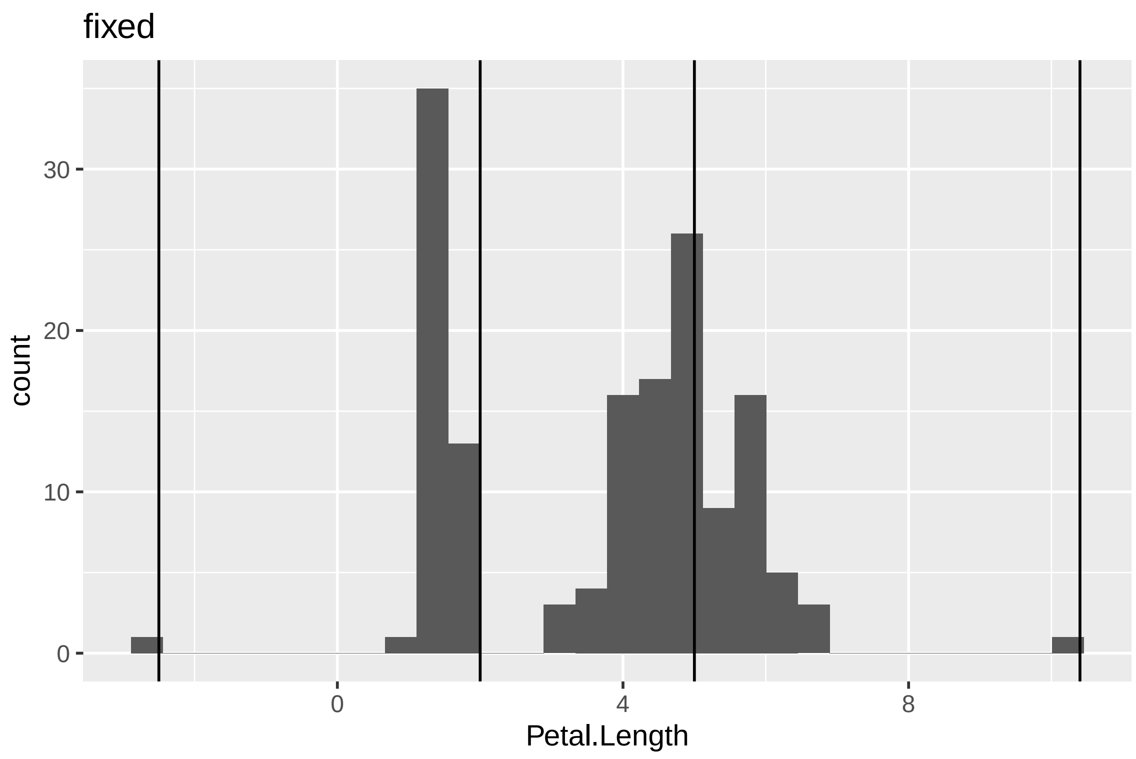


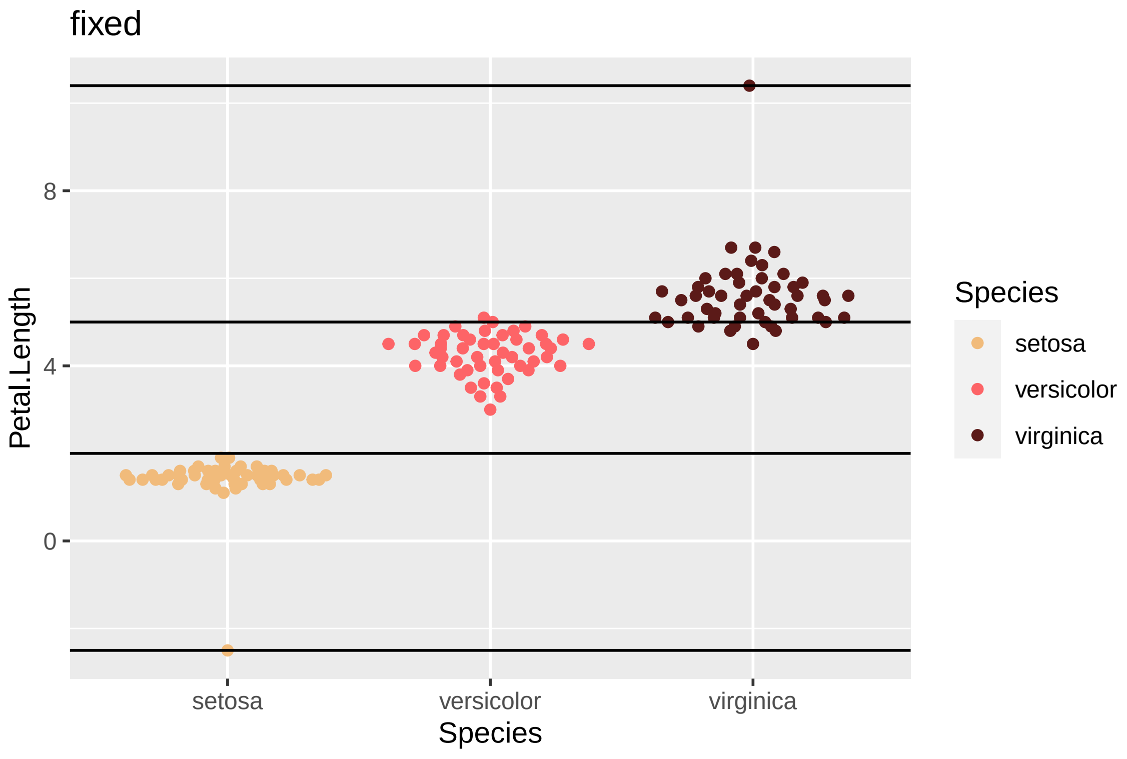
```
## Cases in matched pairs: 95.33 %
```





```
## Cases in matched pairs: 93.33 %
```

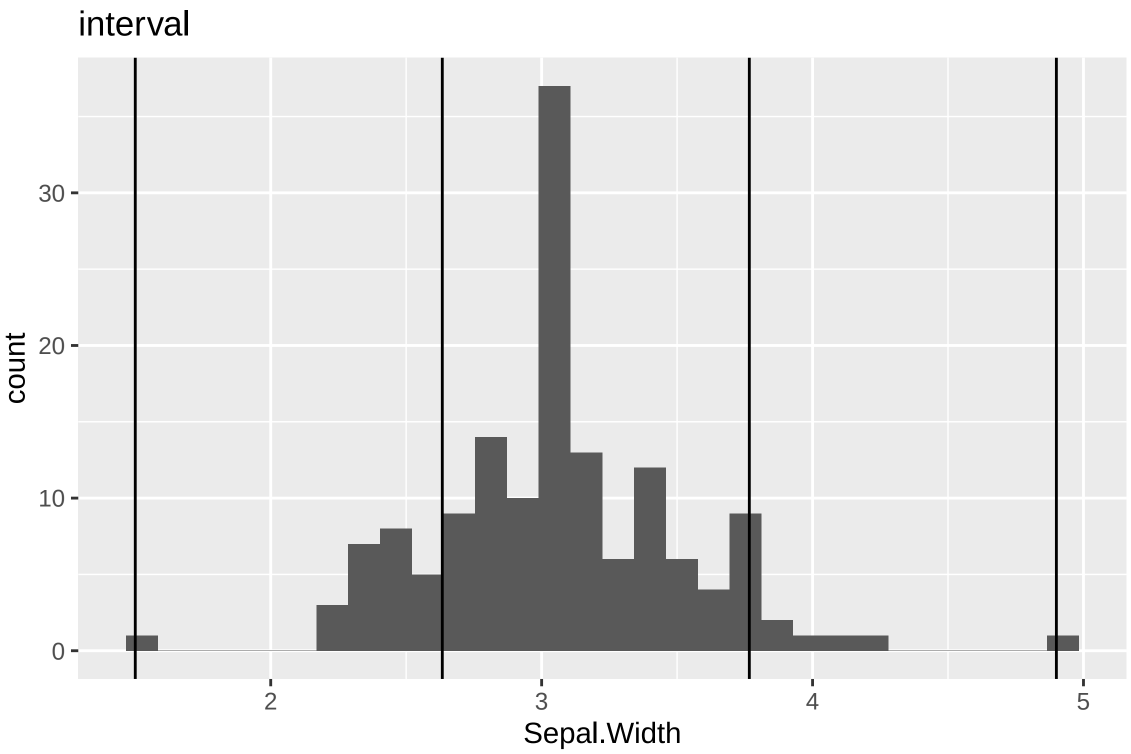




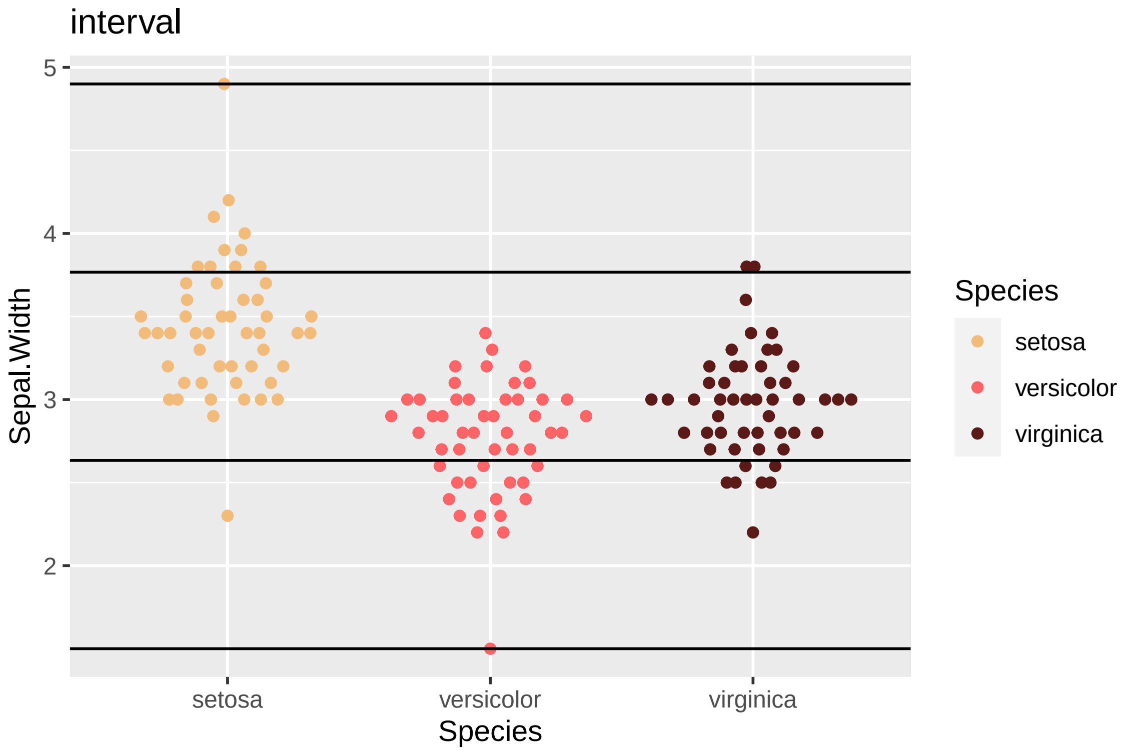
```
## Cases in matched pairs: 94.67 %
```

```
iris$Sepal.Width[which.min(iris$Sepal.Width)] <- min(iris$Sepal.Width) - IQR(iris$Sepal.Width)
iris$Sepal.Width[which.max(iris$Sepal.Width)] <- max(iris$Sepal.Width) + IQR(iris$Sepal.Width)
intervals <- c(min(iris$Sepal.Width), 2.5, 3, max(iris$Sepal.Width))
```

```r
for (method in c("interval", "frequency", "cluster", "fixed")) {
  sepal.width.discretized <- if (method != "fixed")
    discretize(iris$Sepal.Width, method=method) else
    discretize(iris$Sepal.Width, method=method, breaks=intervals)
  print(ggplot(iris, aes(Sepal.Width)) +
          geom_histogram() +
          geom_vline(xintercept=attributes(sepal.width.discretized)$"discretized:breaks") +
          ggtitle(method))
  print(ggplot(iris, aes(Species, Sepal.Width)) +
    geom_quasirandom(aes(col=Species)) +
    scale_color_manual(values=wes_palette("GrandBudapest1", 3)) +
    geom_hline(yintercept=attributes(sepal.width.discretized)$"discretized:breaks") +
    ggtitle(method))
  discretized.table <- table(sepal.width.discretized, iris$Species)
  matchClasses(discretized.table)
}
```

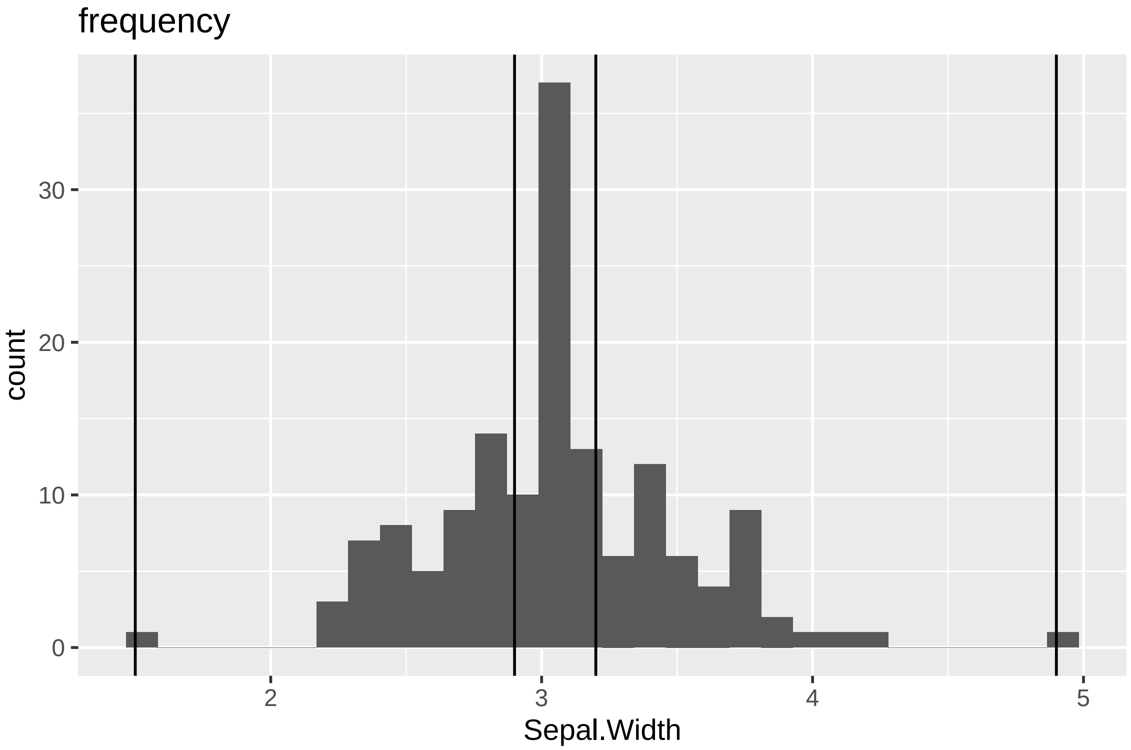


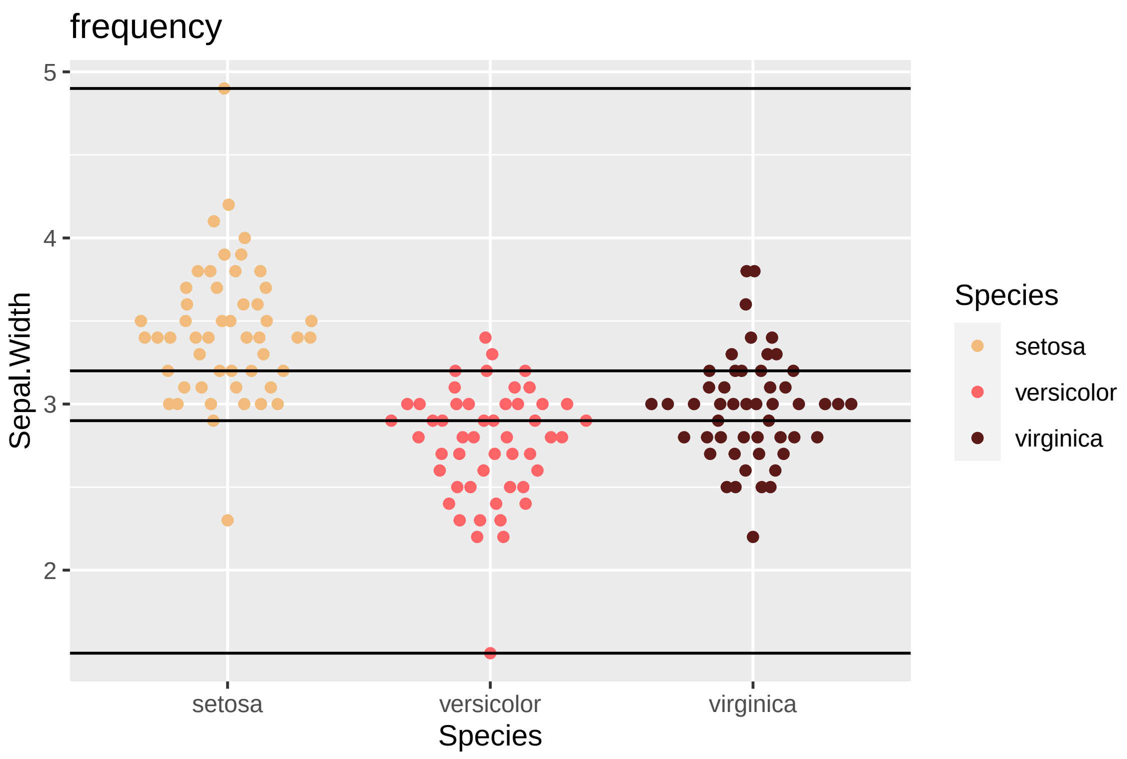
interval



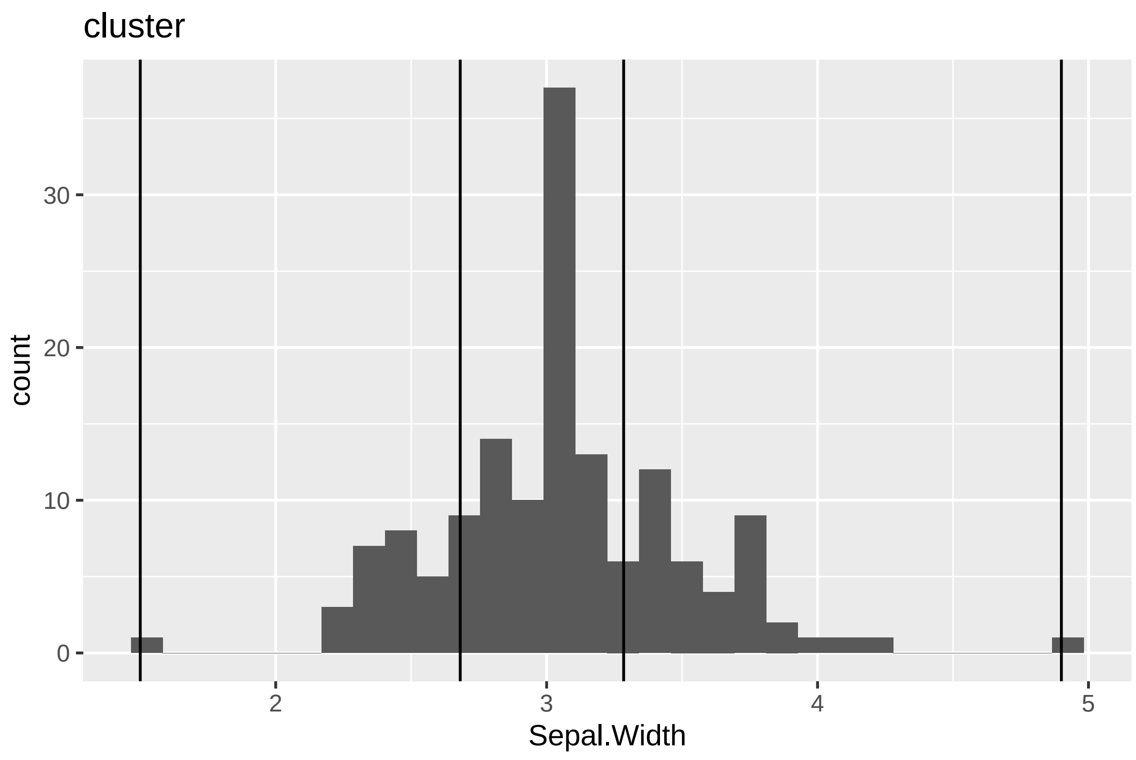
```
## Cases in matched pairs: 44.67 %
```

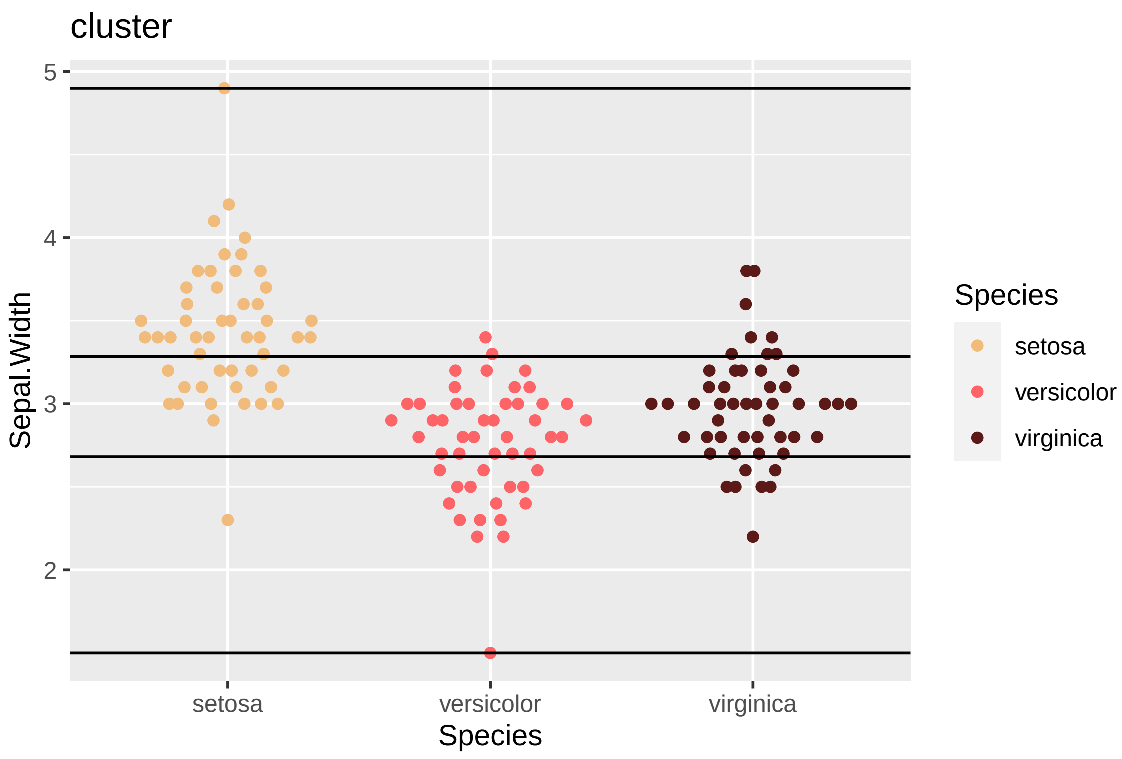
frequency



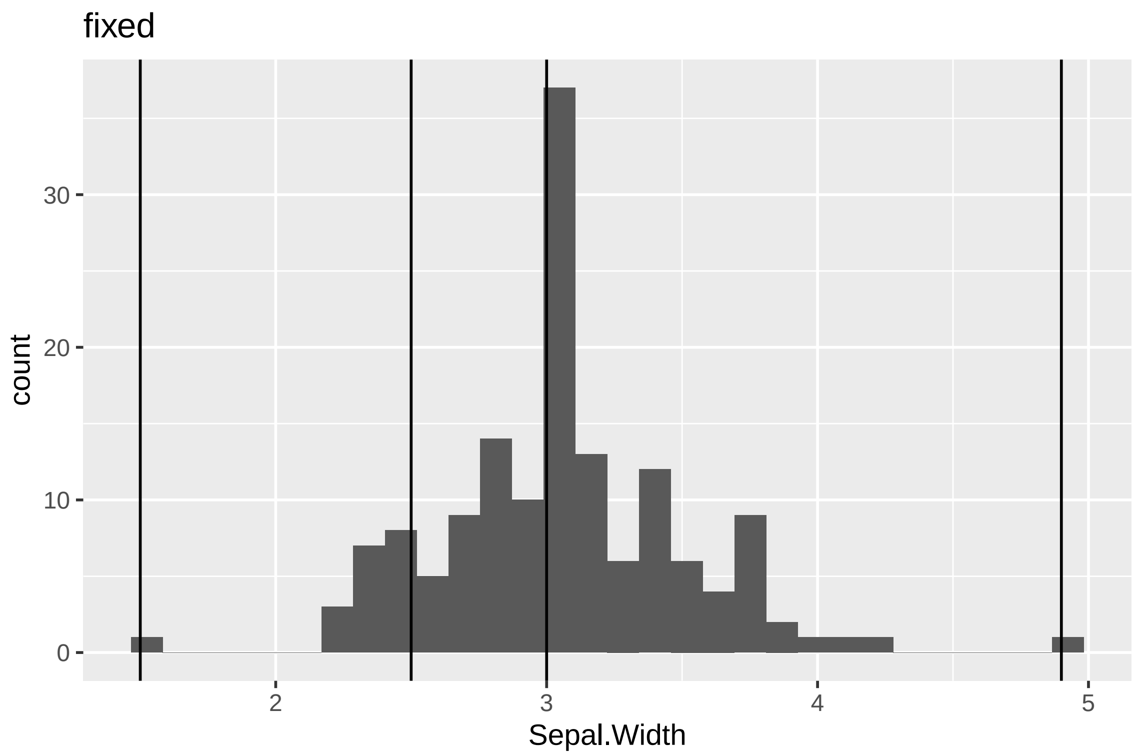


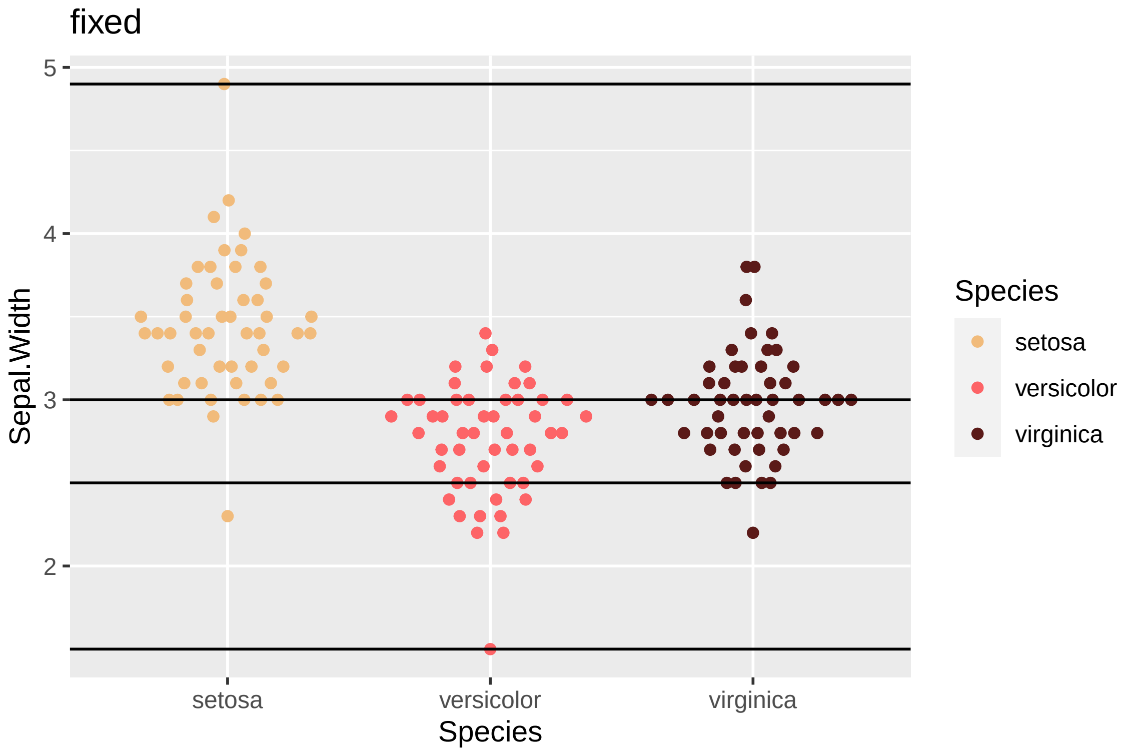
```
## Cases in matched pairs: 55.33 %
```





```
## Cases in matched pairs: 56 %
```



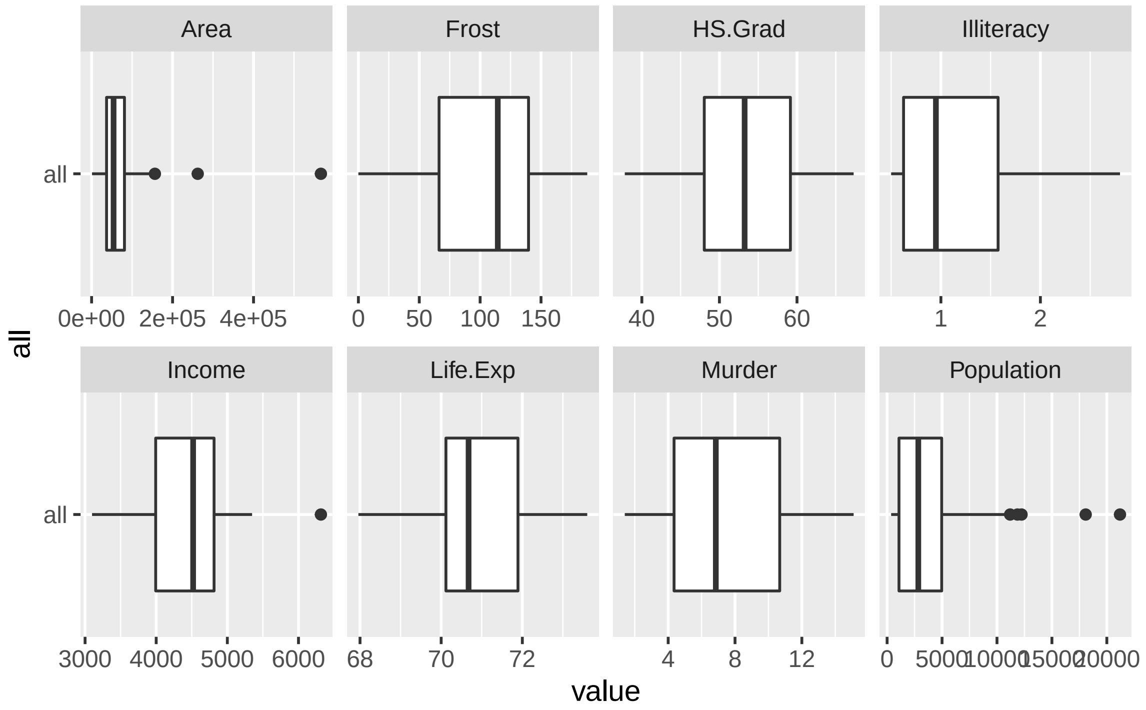


```
## Cases in matched pairs: 54.67 %
```

ZADANIE 2.

```
data(state)
state <- as.data.frame(state.x77)
```

```
plot_boxplot(data.frame(state, all="all"), by="all")
```

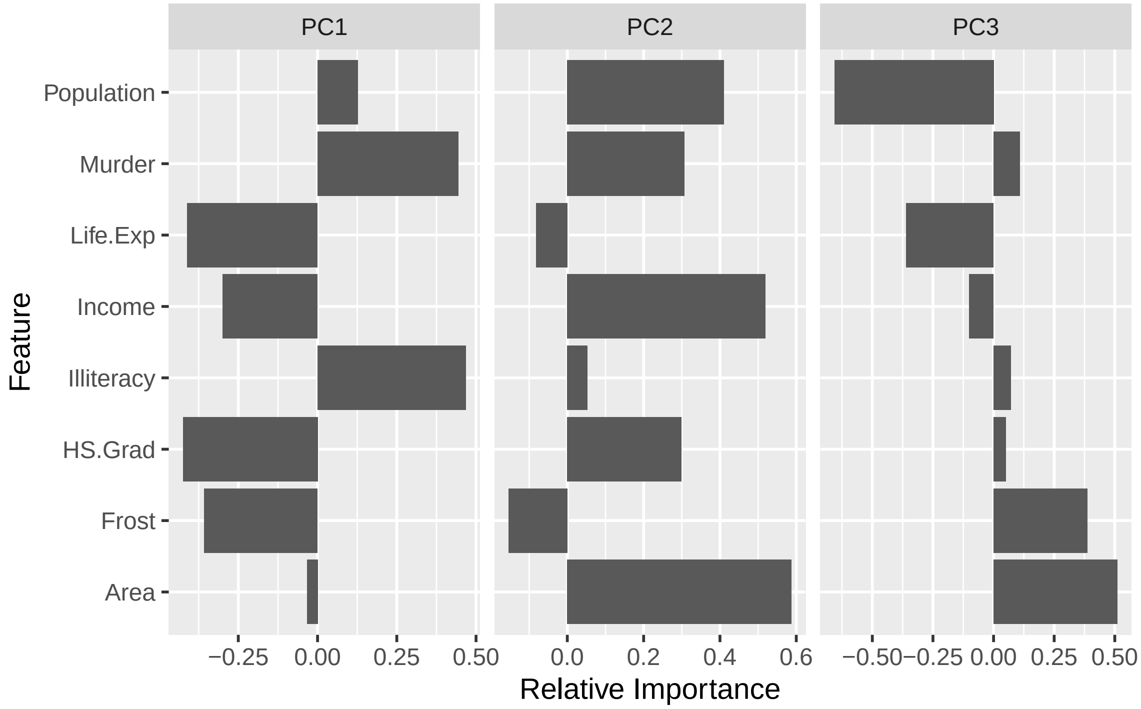


```
sapply(state, function(X) c(sd=sd(X), cv=cv(X)))
```

```
##      Population       Income Illiteracy    Life Exp     Murder    HS Grad       Frost
## sd 4464.491433 614.4699392  0.6095331 1.34239355 3.6915397 8.0769978 51.9808481
## cv    1.051354   0.1385252  0.5209685 0.01893934 0.5003442 0.1520863  0.4976149
##            Area
## sd 85327.29962
## cv     1.20628
```

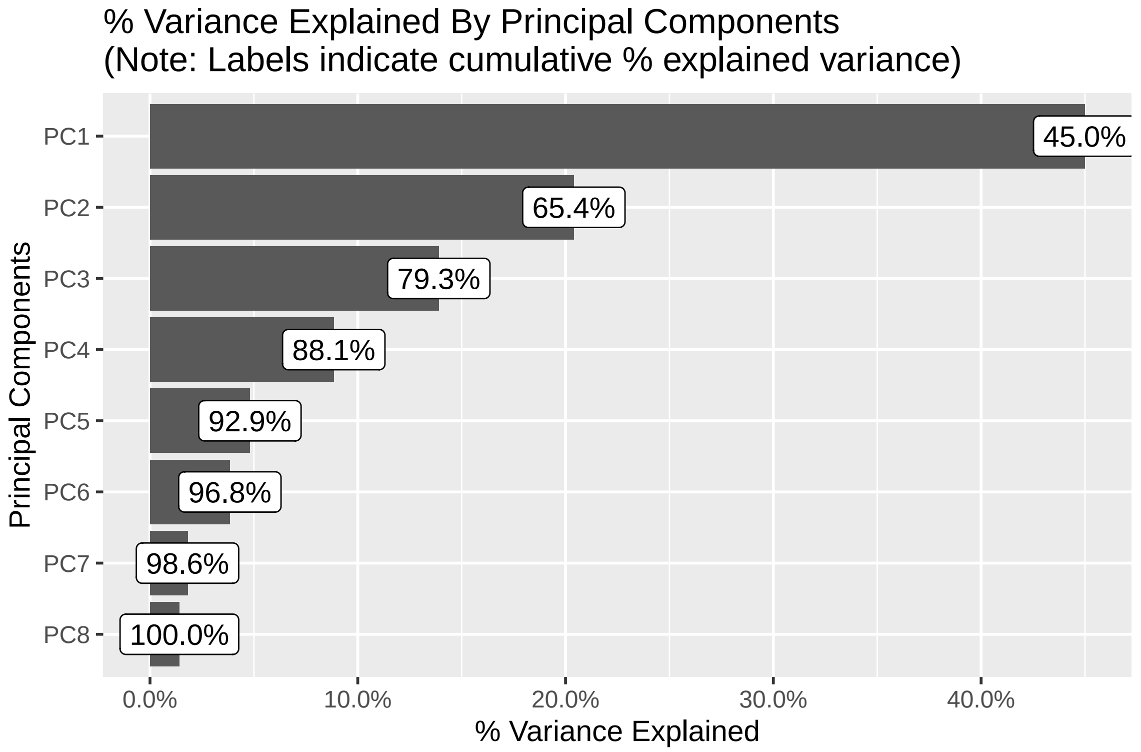
```
p <- plot_prcomp(state, prcomp_args=list(scale=TRUE, center=TRUE), variance_cap=0.8)[2]
```

```
print(p)
```

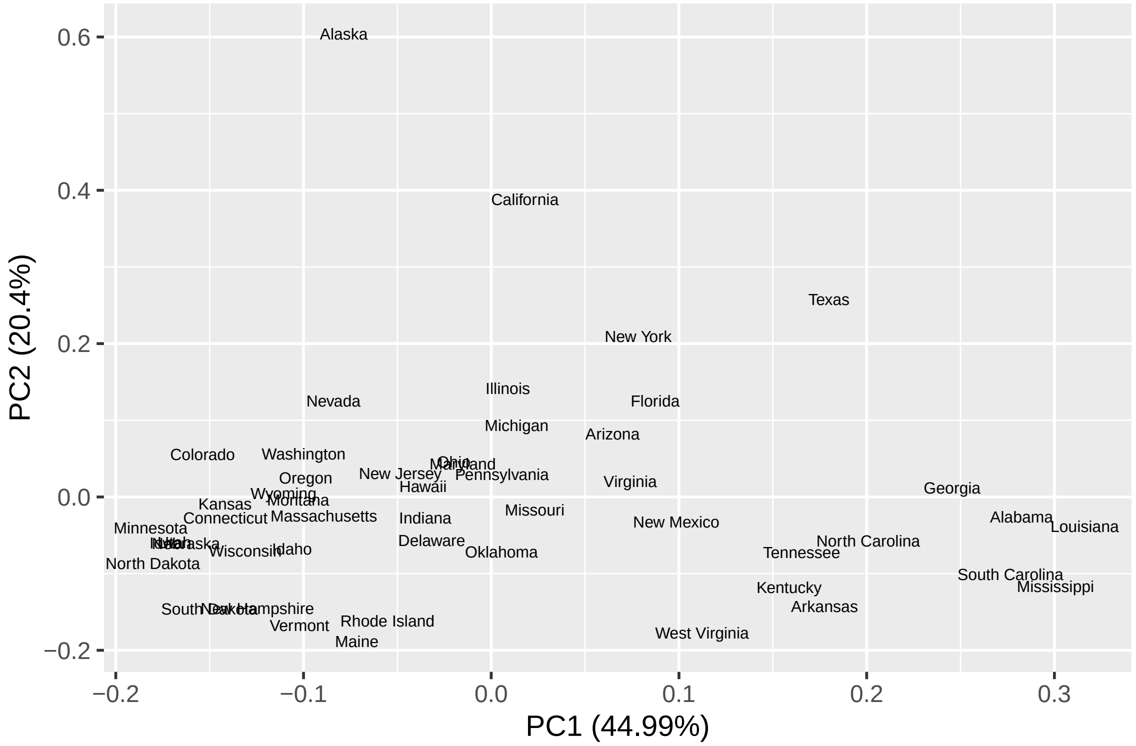


```
p <- plot_prcomp(state, prcomp_args=list(scale=TRUE, center=TRUE), variance_cap=1)[1]
```

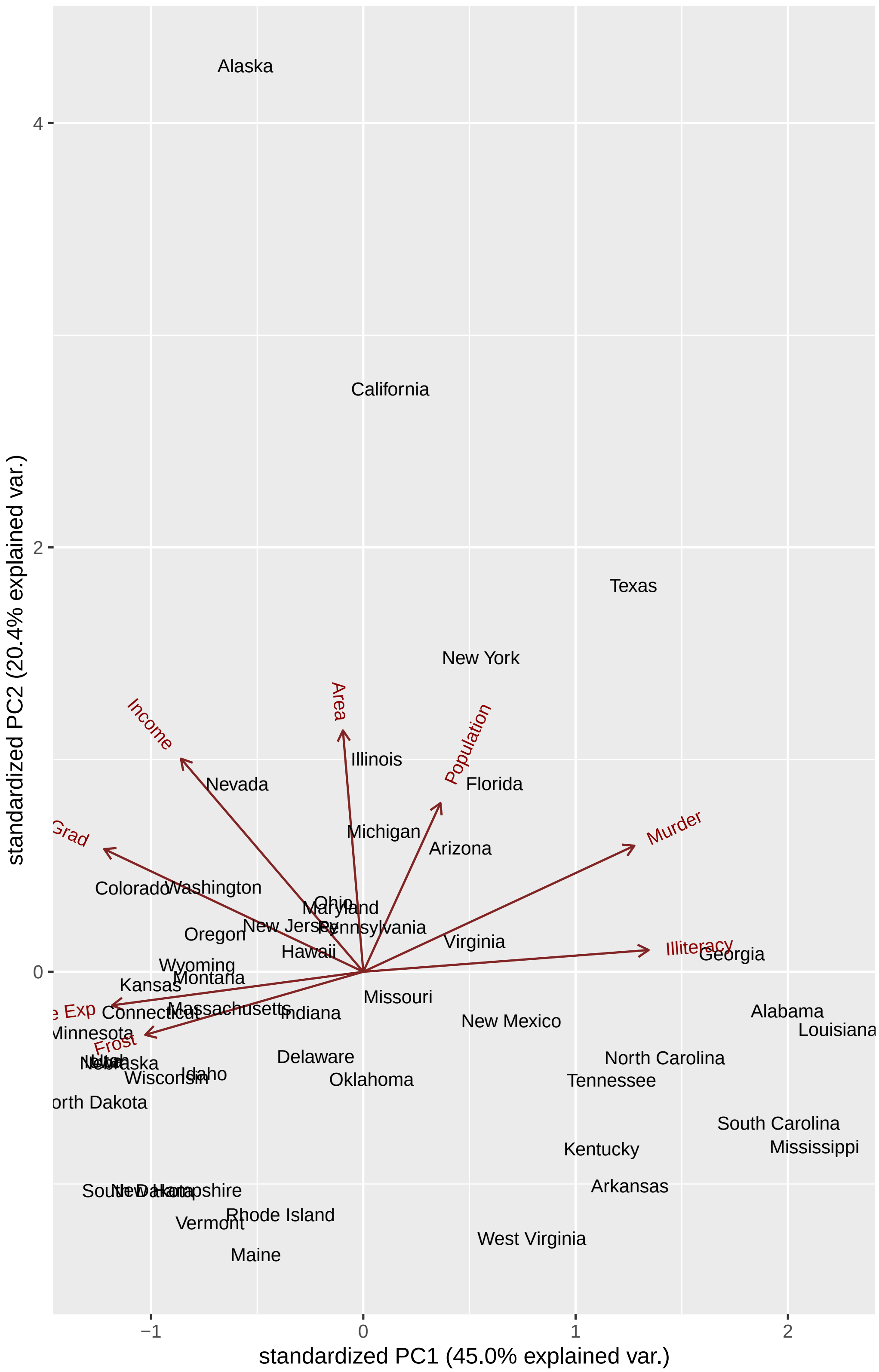
```
print(p)
```



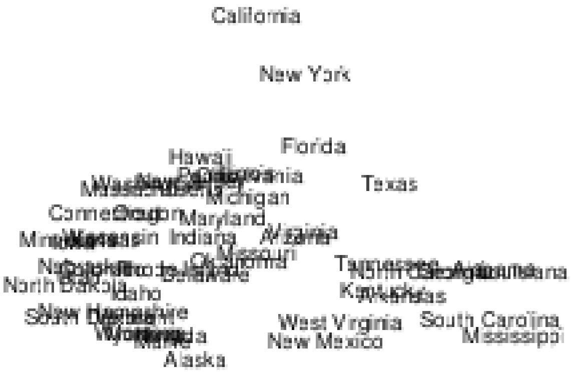
```r
pca <- prcomp(state, center=TRUE, scale=TRUE)
autoplot(pca, shape=FALSE, label.size=2)
```



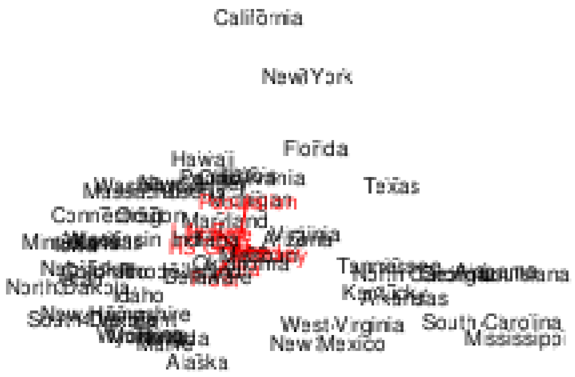
```r
ggbiplot(pca, labels=rownames(state), label.size=2)
```


Alaska
California
Texas
New York
Income
Area
Population
Illinois
Nevada
Florida
Michigan
Murder
Grad
Arizona
Colorado
Washington
Ohio
Maryland
Oregon
New Jersey
Pennsylvania
Virginia
Hawaii
Illiteracy
Georgia
Wyoming
Montana
Kansas
e Exp
Connecticut
Massachusetts
Indiana
Missouri
Alabama
New Mexico
Louisiana
Minnesota
Frost
Delaware
North Carolina
Utah
Iowa
Nebraska
Wisconsin
Idaho
Oklahoma
Tennessee
orth Dakota
South Carolina
Kentucky
Mississippi
Arkansas
South Dakota
New Hampshire
Rhode Island
Vermont
West Virginia
Maine
4
2
0
−1
0
1
2
standardized PC2 (20.4% explained var.)
standardized PC1 (45.0% explained var.)

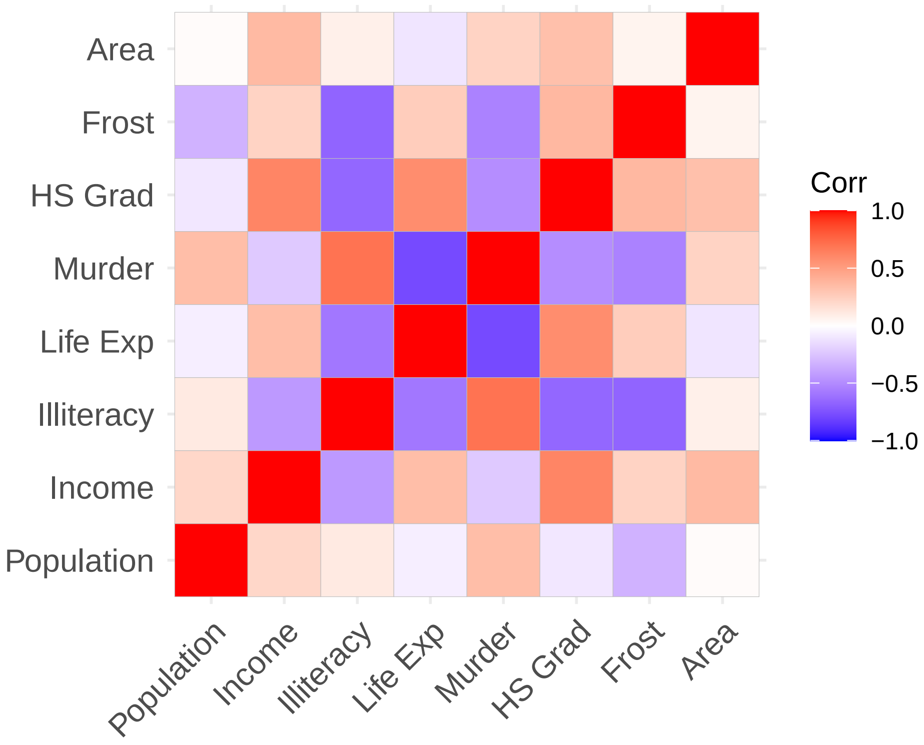
```r
pca <- princomp(state, center=TRUE, scale=TRUE, scores=TRUE, cor=TRUE)
text3d(pca$scores, texts=rownames(state), cex=.5)
rglwidget()
```



```r
text3d(pca$scores, texts=rownames(iris), cex=.5)
text3d(pca$loadings, texts=rownames(pca$loadings), col="red", cex=.5)
coords <- NULL
for (i in 1:nrow(pca$loadings)) {
  coords <- rbind(coords, rbind(c(0,0,0),pca$loadings[i,]))
}
lines3d(coords, col="red", lwd=1)
rglwidget()
```



```
ggcorrplot(cor(state))
```


Area
Frost
HS Grad
Murder
Life Exp
Illiteracy
Income
Population
Population
Income
Illiteracy
Life Exp
Murder
HS Grad
Frost
Area
Corr