Categorical Iris Analysis

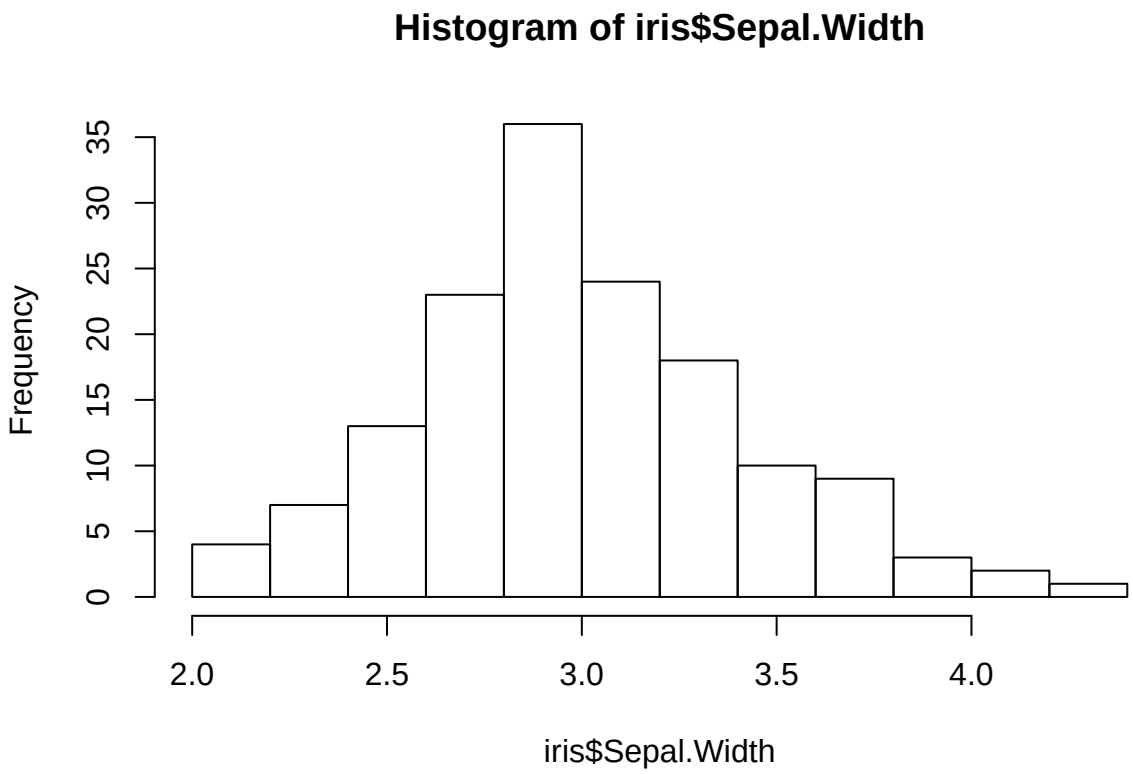
```
data(iris)
```

1. Decide how to discretize each variable.
2. Create a factorized version of the data using binning
3. Look at the relationships between variables.
4. How do we discretize?

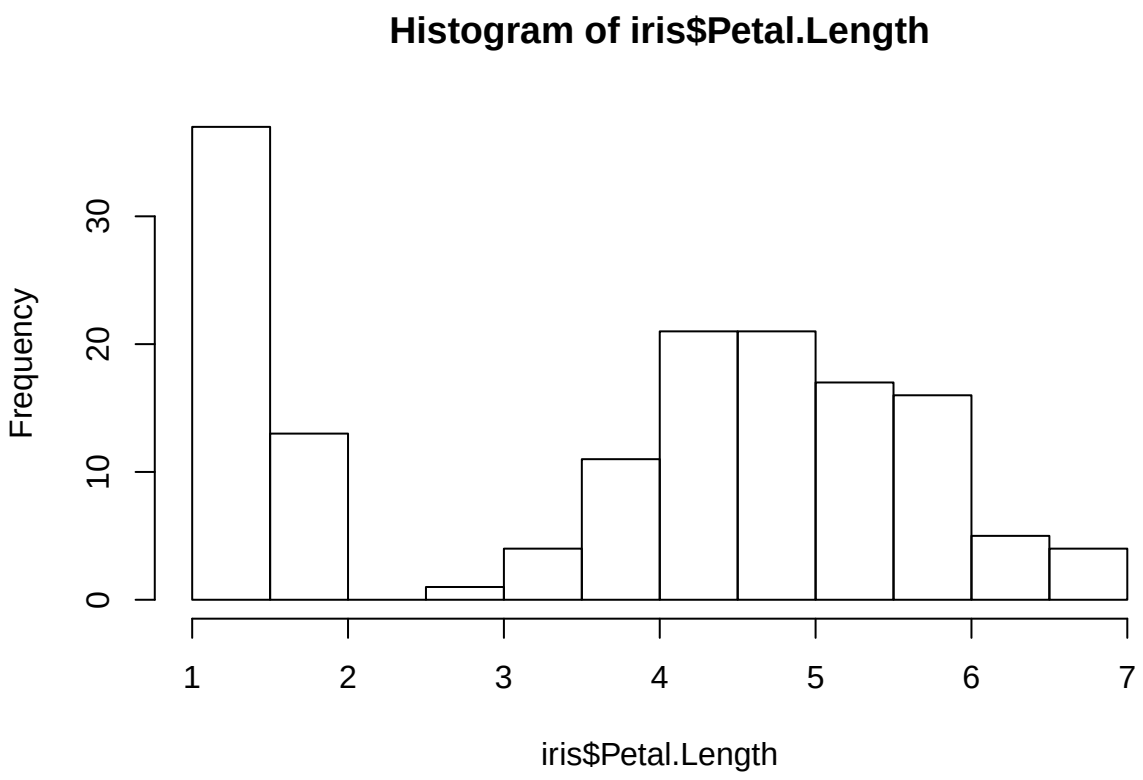
Histograms can tell us the shape of the data.

```
hist(iris$Sepal.Length)
```

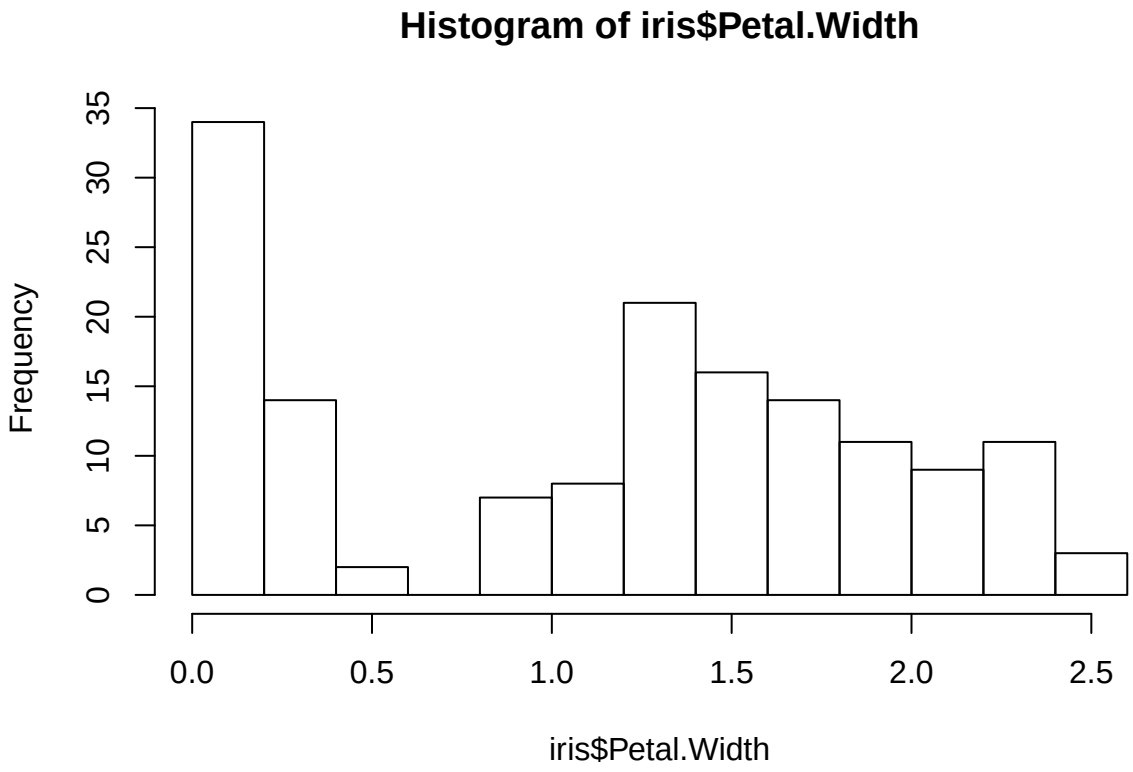

```
hist(iris$Sepal.Width)
```




```
hist(iris$Petal.Length)
```




```
hist(iris$Petal.Width)
```



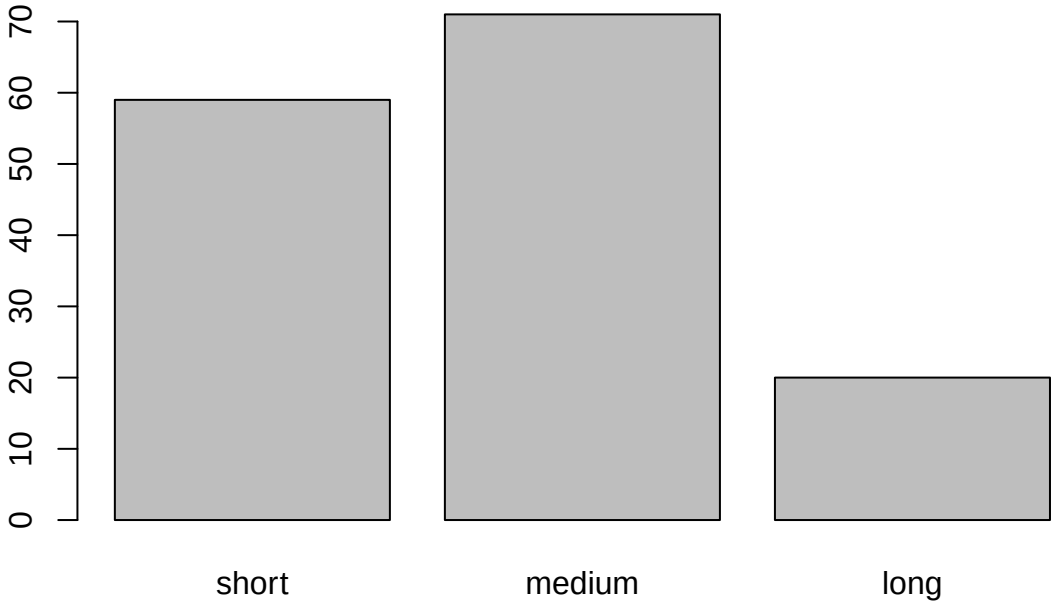
2. Create a factorized version of the data using binning

In R, the "cut" function turns continuous data into factors.

```
Sepal.Length <- cut(iris$Sepal.Length, 3, labels=c('short', 'medium', 'long'))
head(Sepal.Length)
```

```
## [1] short short short short short short
## Levels: short medium long
```

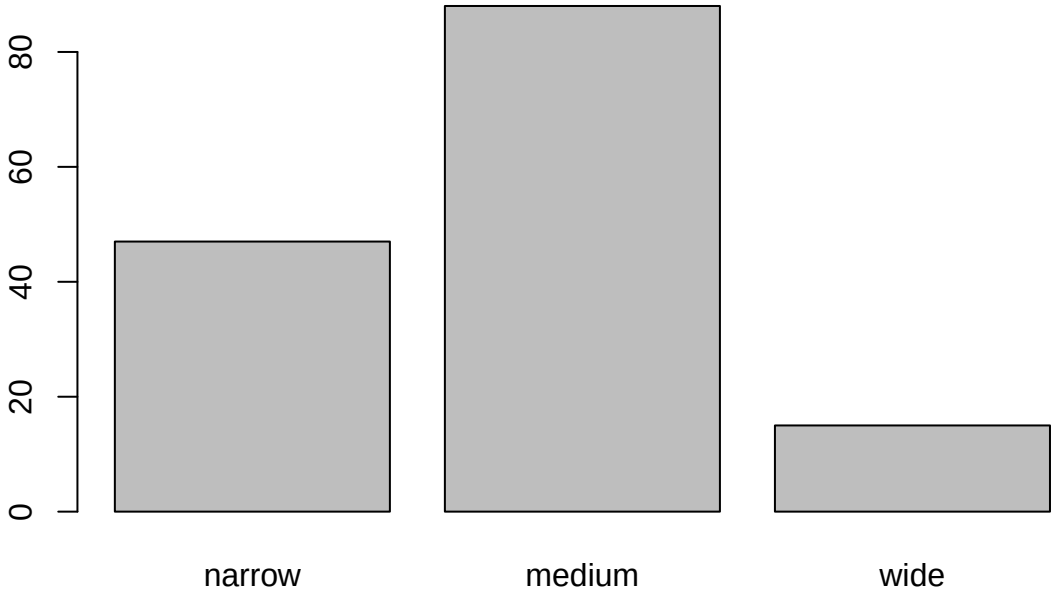
```
barplot(table(Sepal.Length))
```



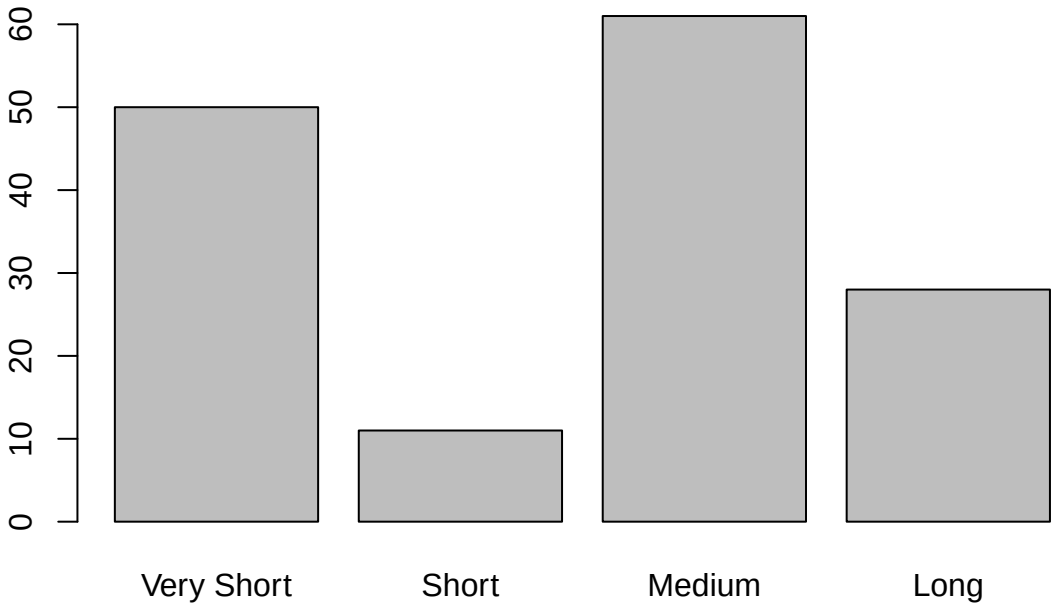
```r
table(Sepal.Length)
```

```
## Sepal.Length
##  short medium   long
##     59     71     20
```

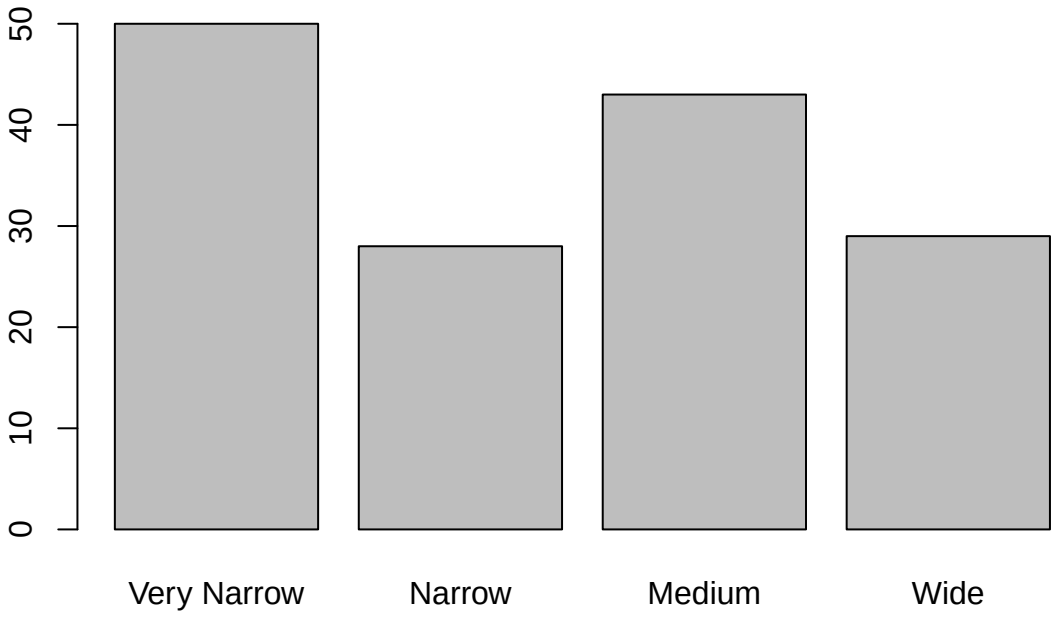
```r
Sepal.Width <- cut(iris$Sepal.Width, 3, labels=c("narrow", "medium", "wide"))
barplot(table(Sepal.Width))
```




```r
Petal.Length <- cut(iris$Petal.Length, 4, labels=c('Very Short', 'Short', 'Medium', 'Long'))
barplot(table(Petal.Length))
```




```r
Petal.Width <- cut(iris$Petal.Width, 4, labels=c('Very Narrow', 'Narrow', 'Medium', 'Wide'))
barplot(table(Petal.Width))
```



Let's produce a categorical version of the iris data frame

```
Species <- iris$Species
ciris <- data.frame(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, Species)
head(ciris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1        short      medium   Very Short Very Narrow  setosa
## 2        short      medium   Very Short Very Narrow  setosa
## 3        short      medium   Very Short Very Narrow  setosa
## 4        short      medium   Very Short Very Narrow  setosa
## 5        short      medium   Very Short Very Narrow  setosa
## 6        short        wide   Very Short Very Narrow  setosa
```

3. Look at the relationships between variables.

Table of 2 way counts

```
sl_sw <- table(Sepal.Length, Sepal.Width)
sl_sw
```

```
##              Sepal.Width
## Sepal.Length narrow medium wide
##       short      12     37   10
##       medium     31     37    3
##       long        4     14    2
```

```
sl_sw <- addmargins(sl_sw)
sl_sw
```

```
##              Sepal.Width
## Sepal.Length narrow medium wide Sum
##       short      12     37   10  59
##       medium     31     37    3  71
##       long        4     14    2  20
##       Sum        47     88   15 150
```

Build a contingency table

```
prop.table(sl_sw[1:3,1:3])
```

```
##              Sepal.Width
## Sepal.Length     narrow     medium       wide
##       short  0.08000000 0.24666667 0.06666667
##       medium 0.20666667 0.24666667 0.02000000
##       long   0.02666667 0.09333333 0.01333333
```

x^2 test

```
chisq.test(ciris$Sepal.Length, ciris$Sepal.Width)
```

```
## Warning in chisq.test(ciris$Sepal.Length, ciris$Sepal.Width): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  ciris$Sepal.Length and ciris$Sepal.Width
## X-squared = 12.879, df = 4, p-value = 0.01188
```

Let's group these commands into a function

```
cat_test <- function(x, y) {
  print(addmargins(table(x,y)))
  print(prop.table(table(x,y)))
  print(chisq.test(x,y))
}
```

```
cat_test(ciris$Sepal.Length, ciris$Sepal.Width)
```

```
##         y
## x        narrow medium wide Sum
##   short      12     37   10  59
##   medium     31     37    3  71
##   long        4     14    2  20
##   Sum        47     88   15 150
##         y
## x            narrow     medium       wide
##   short  0.08000000 0.24666667 0.06666667
##   medium 0.20666667 0.24666667 0.02000000
##   long   0.02666667 0.09333333 0.01333333
```

```
## Warning in chisq.test(x, y): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  x and y
## X-squared = 12.879, df = 4, p-value = 0.01188
```

```
cat_test(ciris$Petal.Length, ciris$Petal.Width)
```

```
##             y
## x            Very Narrow Narrow Medium Wide Sum
##   Very Short          50      0      0    0  50
##   Short               0     10      1    0  11
##   Medium              0     18     33   10  61
##   Long                0      0      9   19  28
```

```
##   Sum                50     28     43   29 150
##             y
## x            Very Narrow      Narrow      Medium        Wide
##   Very Short 0.333333333 0.000000000 0.000000000 0.000000000
##   Short      0.000000000 0.066666667 0.006666667 0.000000000
##   Medium     0.000000000 0.120000000 0.220000000 0.066666667
##   Long       0.000000000 0.000000000 0.060000000 0.126666667
```

```
## Warning in chisq.test(x, y): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  x and y
## X-squared = 225.01, df = 9, p-value < 2.2e-16
```

```
cat_test(ciris$Sepal.Length, ciris$Petal.Length)
```

```
##         y
## x        Very Short Short Medium Long Sum
##   short          47     7      5    0  59
##   medium          3     4     51   13  71
##   long            0     0      5   15  20
##   Sum            50    11     61   28 150
##         y
## x        Very Short      Short     Medium       Long
##   short  0.31333333 0.04666667 0.03333333 0.00000000
##   medium 0.02000000 0.02666667 0.34000000 0.08666667
##   long   0.00000000 0.00000000 0.03333333 0.10000000
```

```
## Warning in chisq.test(x, y): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  x and y
## X-squared = 144.32, df = 6, p-value < 2.2e-16
```

```
cat_test(Sepal.Length, Petal.Width)
```

```
##         y
## x        Very Narrow Narrow Medium Wide Sum
##   short           47      9      3    0  59
##   medium           3     19     33   16  71
##   long             0      0      7   13  20
##   Sum             50     28     43   29 150
##         y
## x        Very Narrow     Narrow     Medium       Wide
##   short   0.31333333 0.06000000 0.02000000 0.00000000
##   medium  0.02000000 0.12666667 0.22000000 0.10666667
##   long    0.00000000 0.00000000 0.04666667 0.08666667
```

```
## Warning in chisq.test(x, y): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  x and y
```

```
## X-squared = 122.24, df = 6, p-value < 2.2e-16
```

```
cat_test(ciris$Sepal.Length, ciris$Species)
```

```
##         y
## x        setosa versicolor virginica Sum
##   short      47         11         1  59
##   medium      3         36        32  71
##   long        0          3        17  20
##   Sum        50         50        50 150
##         y
## x             setosa  versicolor   virginica
##   short  0.313333333 0.073333333 0.006666667
##   medium 0.020000000 0.240000000 0.213333333
##   long   0.000000000 0.020000000 0.113333333
##
##  Pearson's Chi-squared test
##
## data:  x and y
## X-squared = 111.63, df = 4, p-value < 2.2e-16
```

Separate the categories of data

```
setosa <- ciris[ciris$Species=='setosa', 1:4]
setosa
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1         short      medium   Very Short Very Narrow
## 2         short      medium   Very Short Very Narrow
## 3         short      medium   Very Short Very Narrow
## 4         short      medium   Very Short Very Narrow
## 5         short      medium   Very Short Very Narrow
## 6         short        wide   Very Short Very Narrow
## 7         short      medium   Very Short Very Narrow
## 8         short      medium   Very Short Very Narrow
## 9         short      medium   Very Short Very Narrow
## 10        short      medium   Very Short Very Narrow
## 11        short        wide   Very Short Very Narrow
## 12        short      medium   Very Short Very Narrow
## 13        short      medium   Very Short Very Narrow
## 14        short      medium   Very Short Very Narrow
## 15       medium        wide   Very Short Very Narrow
## 16       medium        wide   Very Short Very Narrow
## 17        short        wide   Very Short Very Narrow
## 18        short      medium   Very Short Very Narrow
## 19       medium        wide   Very Short Very Narrow
## 20        short        wide   Very Short Very Narrow
## 21        short      medium   Very Short Very Narrow
## 22        short        wide   Very Short Very Narrow
## 23        short      medium   Very Short Very Narrow
## 24        short      medium   Very Short Very Narrow
## 25        short      medium   Very Short Very Narrow
## 26        short      medium   Very Short Very Narrow
## 27        short      medium   Very Short Very Narrow
## 28        short      medium   Very Short Very Narrow
## 29        short      medium   Very Short Very Narrow
```

```
## 30        short      medium   Very Short Very Narrow
## 31        short      medium   Very Short Very Narrow
## 32        short      medium   Very Short Very Narrow
## 33        short        wide   Very Short Very Narrow
## 34        short        wide   Very Short Very Narrow
## 35        short      medium   Very Short Very Narrow
## 36        short      medium   Very Short Very Narrow
## 37        short      medium   Very Short Very Narrow
## 38        short      medium   Very Short Very Narrow
## 39        short      medium   Very Short Very Narrow
## 40        short      medium   Very Short Very Narrow
## 41        short      medium   Very Short Very Narrow
## 42        short      narrow   Very Short Very Narrow
## 43        short      medium   Very Short Very Narrow
## 44        short      medium   Very Short Very Narrow
## 45        short        wide   Very Short Very Narrow
## 46        short      medium   Very Short Very Narrow
## 47        short        wide   Very Short Very Narrow
## 48        short      medium   Very Short Very Narrow
## 49        short        wide   Very Short Very Narrow
## 50        short      medium   Very Short Very Narrow
```

```
cat_test(setosa$Sepal.Length, setosa$Sepal.Width)
```

```
##          y
## x         narrow medium wide Sum
##   short        1     36   10  47
##   medium       0      0    3   3
##   long         0      0    0   0
##   Sum          1     36   13  50
##          y
## x         narrow medium wide
##   short    0.02   0.72 0.20
##   medium   0.00   0.00 0.06
##   long     0.00   0.00 0.00
```

```
## Warning in chisq.test(x, y): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  x and y
## X-squared = 9.0835, df = 2, p-value = 0.01065
```

```
#cat_test(setosa$Petal.Length, setosa$Petal.Width)
#That fails because there is only one level!
```

```
versicolor <- ciris[ciris$Species=='versicolor', 1:4]
```

Sepal

```
cat_test(versicolor$Sepal.Length, versicolor$Sepal.Width)
```

```
##          y
## x         narrow medium wide Sum
##   short       10      1    0  11
##   medium      16     20    0  36
```

```
##   long        1      2    0   3
##   Sum        27     23    0  50
##         y
## x        narrow medium wide
##   short    0.20   0.02 0.00
##   medium   0.32   0.40 0.00
##   long     0.02   0.04 0.00
```

```
## Warning in chisq.test(x, y): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  x and y
## X-squared = 7.8718, df = 2, p-value = 0.01953
```

```
cat_test(versicolor$Petal.Length, versicolor$Petal.Width)
```

```
##             y
## x            Very Narrow Narrow Medium Wide Sum
##   Very Short           0      0      0    0   0
##   Short                0     10      1    0  11
##   Medium               0     18     21    0  39
##   Long                 0      0      0    0   0
##   Sum                  0     28     22    0  50
##             y
## x            Very Narrow Narrow Medium Wide
##   Very Short        0.00   0.00   0.00 0.00
##   Short             0.00   0.20   0.02 0.00
##   Medium            0.00   0.36   0.42 0.00
##   Long              0.00   0.00   0.00 0.00
```

```
## Warning in chisq.test(x, y): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  x and y
## X-squared = 5.2767, df = 1, p-value = 0.02161
```

```
virginica <- ciris[ciris$Species=='virginica',1:4]
```

```
cat_test(virginica$Sepal.Length, virginica$Sepal.Width)
```

```
##         y
## x        narrow medium wide Sum
##   short       1      0    0   1
##   medium     15     17    0  32
##   long        3     12    2  17
##   Sum        19     29    2  50
##         y
## x        narrow medium wide
##   short    0.02   0.00 0.00
##   medium   0.30   0.34 0.00
##   long     0.06   0.24 0.04
```

```
## Warning in chisq.test(x, y): Chi-squared approximation may be incorrect
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  x and y
## X-squared = 8.586, df = 4, p-value = 0.07232
```

```
cat_test(virginica$Petal.Length, virginica$Petal.Width)
```

```
##             y
## x            Very Narrow Narrow Medium Wide Sum
##   Very Short           0      0      0    0   0
##   Short                0      0      0    0   0
##   Medium               0      0     12   10  22
##   Long                 0      0      9   19  28
##   Sum                  0      0     21   29  50
##             y
## x            Very Narrow Narrow Medium Wide
##   Very Short        0.00   0.00   0.00 0.00
##   Short             0.00   0.00   0.00 0.00
##   Medium            0.00   0.00   0.24 0.20
##   Long              0.00   0.00   0.18 0.38
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  x and y
## X-squared = 1.7019, df = 1, p-value = 0.192
```

```
ciris$Sepal[ciris$Sepal.Length == 'short' && ciris$Sepal.Width == 'narrow'] = 'sn'
ciris$Sepal
```

```
##   [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [26] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [51] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [76] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [101] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [126] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```