CS544 Module3

Suresh Kalathur

Module3

- Data Description
  - Univariate Data
  - Bivariate Data
  - Multivariate Data

Types of Data

- Qualitative (categorical) data
  - Nominal data
  - Ordinal data
- Quantitative (numerical) data
  - Interval data
    - Measured on a scale of equal-sized units
  - Ratio data
    - Order of magnitude is also important

Categorical Data

- Non-visual representations
  - Tables
    - table(x)
- Visual representations
  - Barplot, Piechart, etc.
  - Examples

Numerical Data

- Measures of center and spread
  - Mean, median, mode
  - Range, variance, standard deviation
- Five number summary
  - fivenum(x) versus summary(x)
- Quantiles
- Z-scores

…Numerical Data

- Graphical representation
  - Barplot, Dotchart
  - Barplot with frequencies
  - Stem plot
  - Histogram
  - Boxplot

Bivariate Data

- Contingency (two-way) tables
  - Summarize bivariate categorical data
  - table(x,y)
- Marginal Distributions of two-way tables
  - margin.table(…)
- Conditional Distributions of two-way tables
  - prop.table(…)

…Graphical

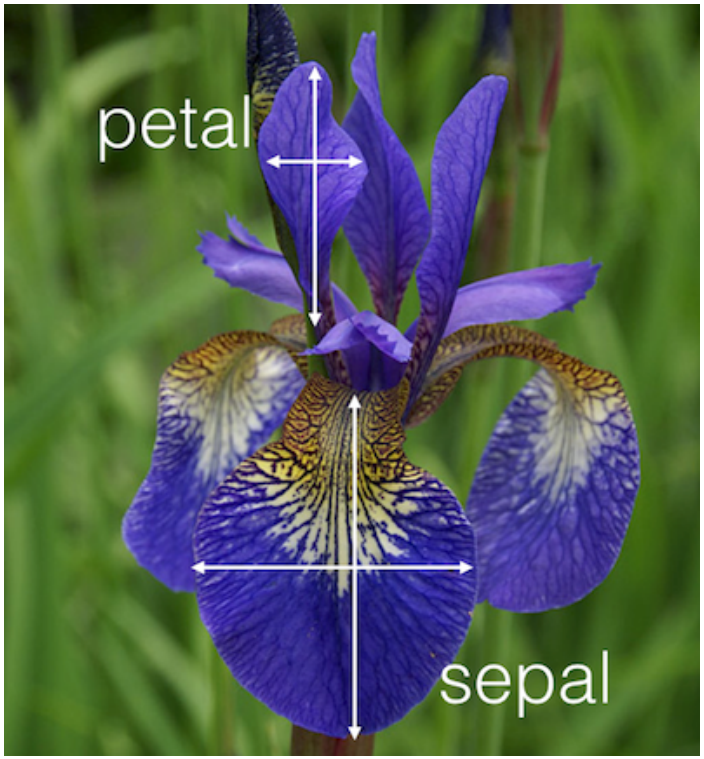
- Mosaic plots
- Bar plots of two-way tables
- Scatter plot
- Pair-wise plot and Correlation
- Other examples
  - IRIS dataset
  - Titanic dataset

Iris Flower Data Set

Iris setosa	Iris versicolor	Iris virginica

… Iris Data Set

- *50 samples from each species*
- *Four features from each sample (in cm.)*
  - *Sepal length*
  - *Sepal width*
  - *Petal length*
  - *Petal width*
- *Class label*

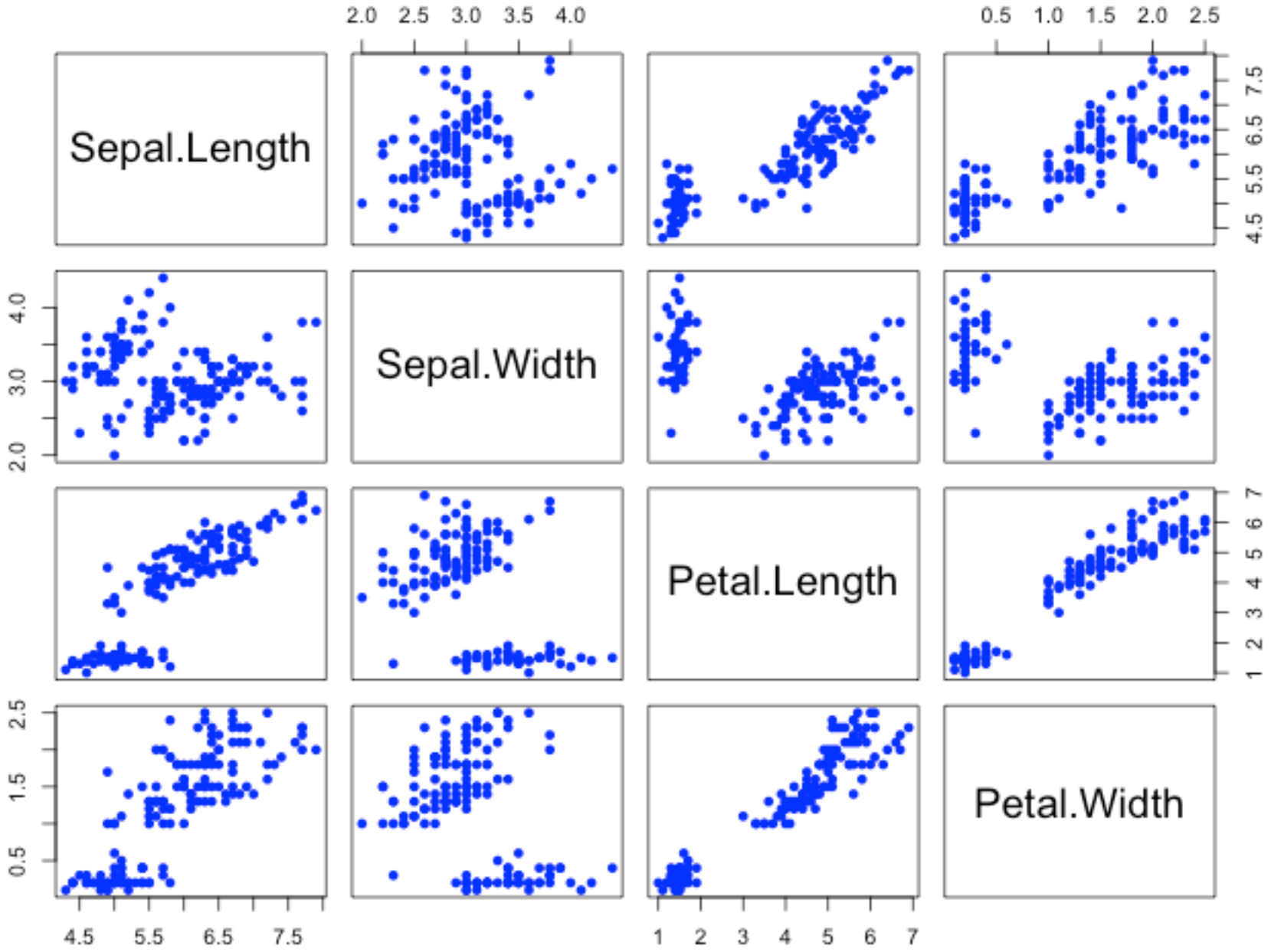


...Iris Dataset

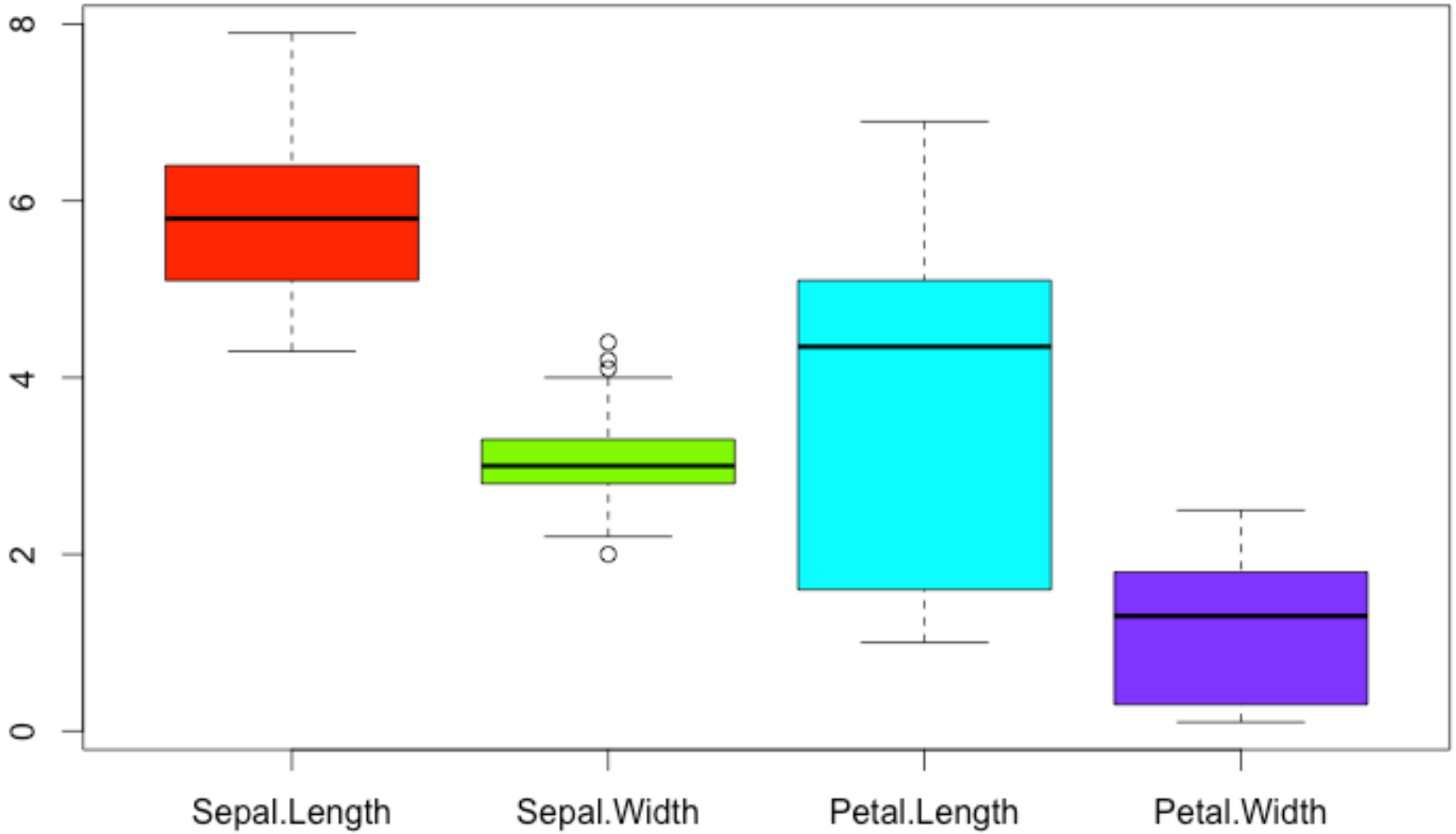
```
> names(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length"
[4] "Petal.Width"  "Species"
> data <- iris[c(1:4)]
>
> summary(data)
  Sepal.Length  Sepal.Width   Petal.Length  Petal.Width
 Min.   :4.3   Min.   :2.0   Min.   :1.0   Min.   :0.1
 1st Qu.:5.1   1st Qu.:2.8   1st Qu.:1.6   1st Qu.:0.3
 Median :5.8   Median :3.0   Median :4.3   Median :1.3
 Mean   :5.8   Mean   :3.1   Mean   :3.8   Mean   :1.2
 3rd Qu.:6.4   3rd Qu.:3.3   3rd Qu.:5.1   3rd Qu.:1.8
 Max.   :7.9   Max.   :4.4   Max.   :6.9   Max.   :2.5
```

- Scatterplot and Correlation matrix

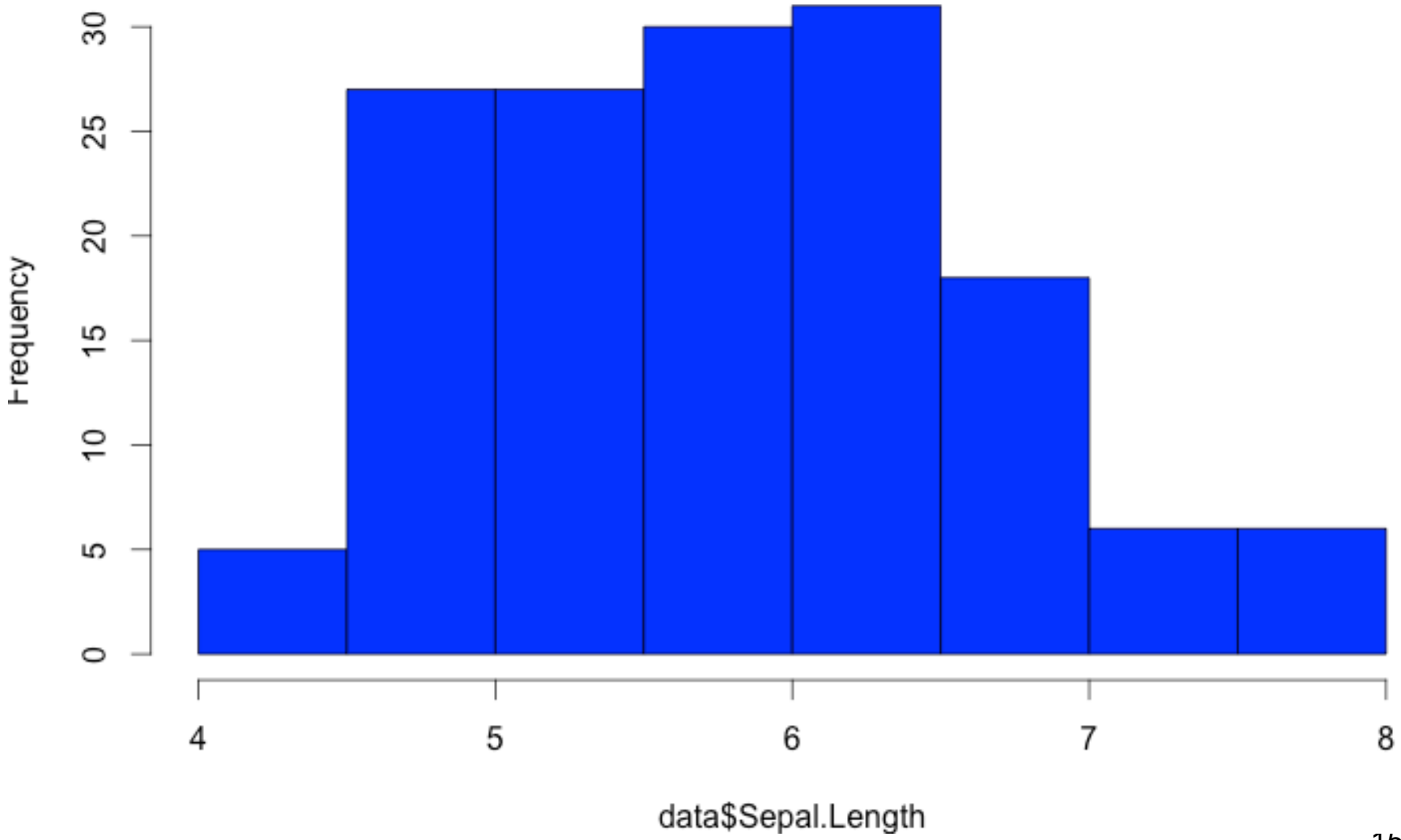
```
> pairs(data, pch=16, col="blue")
> cor(data)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         1.00       -0.12         0.87        0.82
Sepal.Width         -0.12        1.00        -0.43       -0.37
Petal.Length         0.87       -0.43         1.00        0.96
Petal.Width          0.82       -0.37         0.96        1.00
```


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
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

```
> boxplot(data, col=rainbow(4))
```



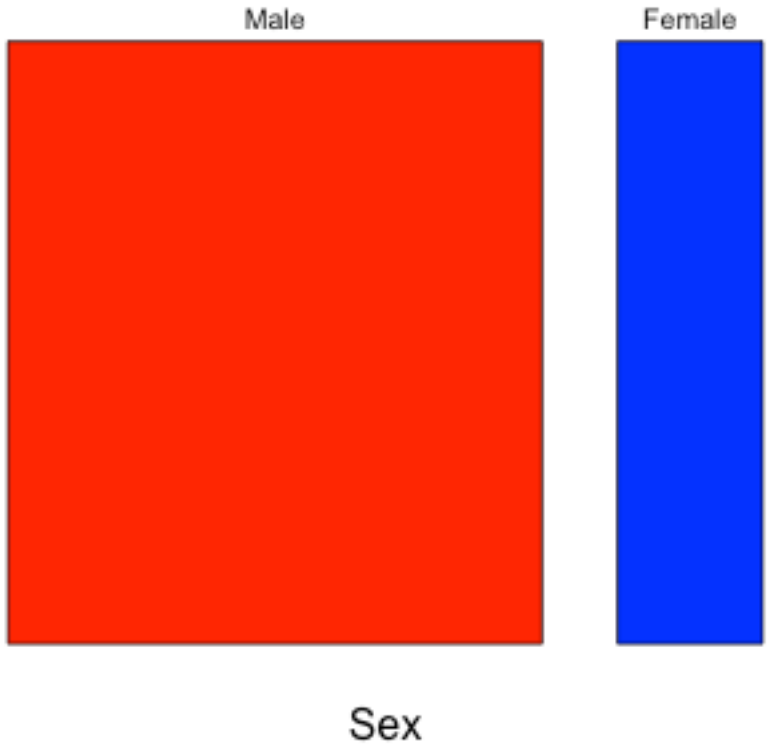
```
hist(data$Sepal.Length, col="blue")
```



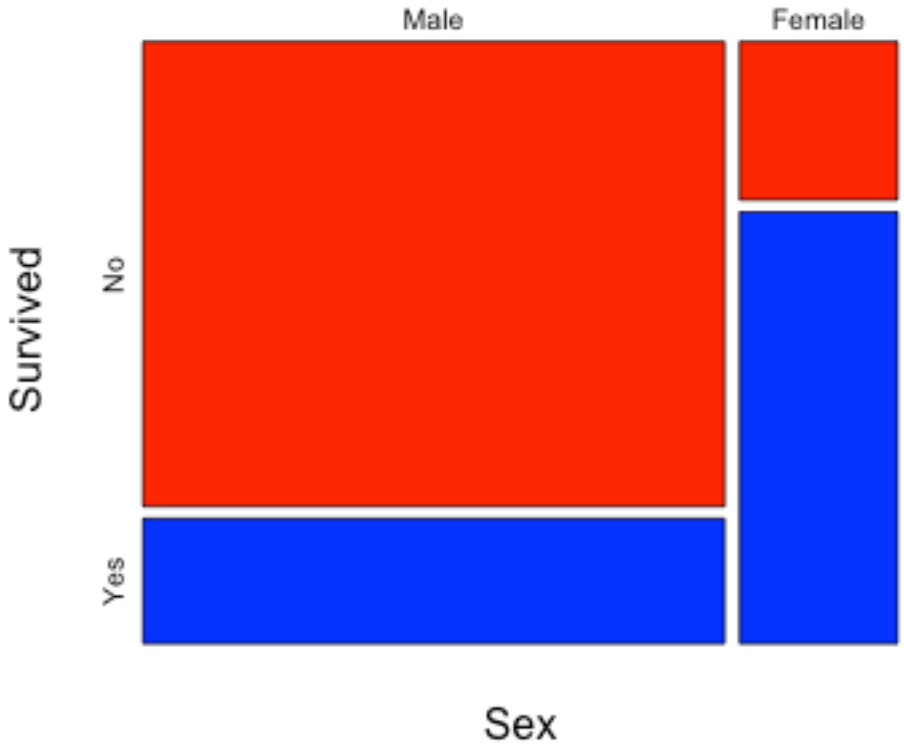
Mosaic Plots

- Titanic Dataset
  - Class – 1st, 2nd, 3rd, Crew
  - Sex – Male, Female
  - Age: Child, Adult
  - Survived: No, Yes

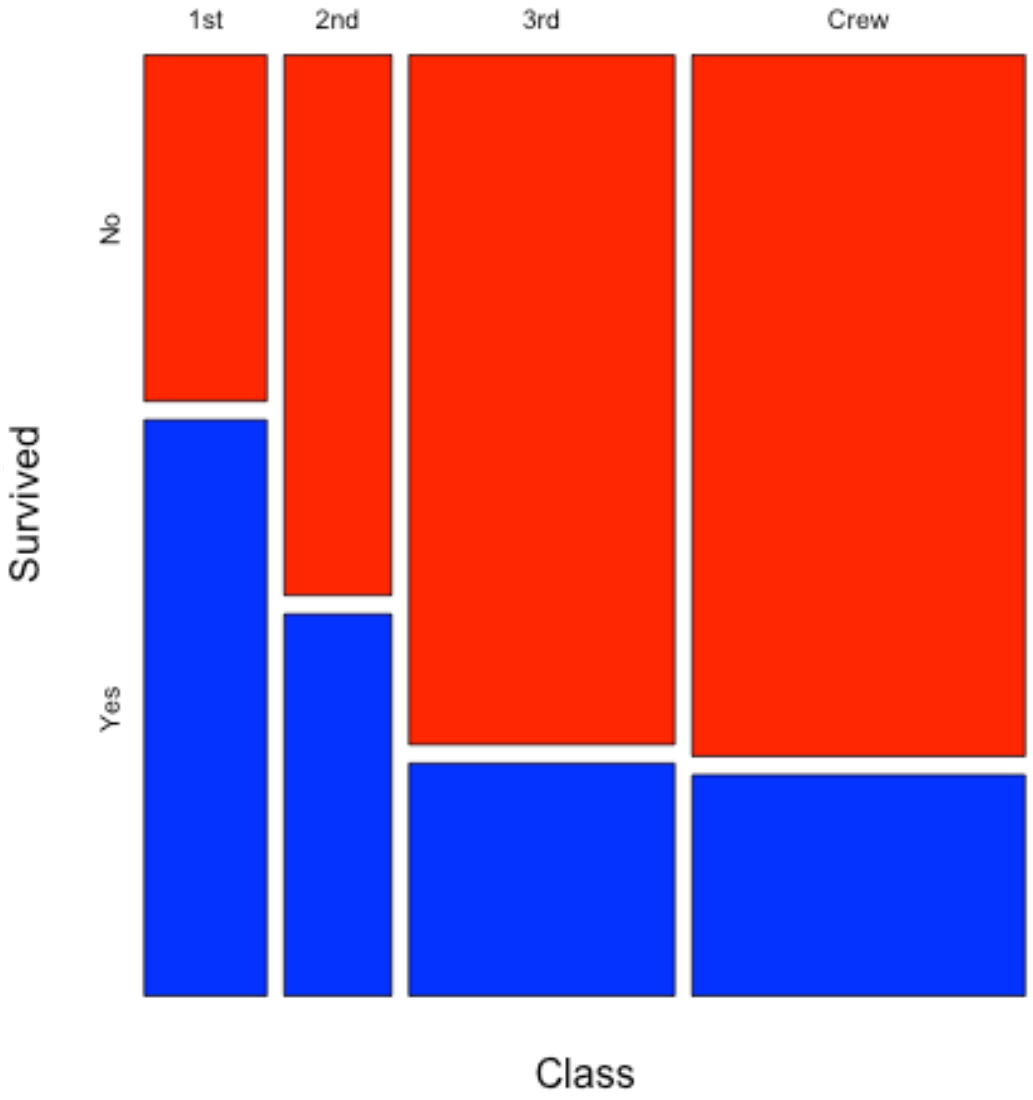
```
> # Sex
> t1 <- margin.table(Titanic, c(2))
> t1
Sex
  Male Female
  1731    470
> mosaicplot(t1, col=c("red", "blue"))
```



```
> # Sex, Survived
> t2 <- margin.table(Titanic, c(2,4))
> t2
        Survived
Sex         No  Yes
  Male    1364  367
  Female   126  344
> mosaicplot(t2, col=c("red", "blue"))
```



```
> # Crew, Survived
> t3 <- margin.table(Titanic, c(1, 4))
> t3
      Survived
Class   No Yes
  1st  122 203
  2nd  167 118
  3rd  528 178
  Crew 673 212
> mosaicplot(t3, col=c("red
```



More R!

- apply
- sweep
- tapply
- split
- lapply
- sapply
- Case Study
  - Lincoln's Gettysburg Address