Data Exploration in R

Data Exploration and Visualization with R

- Data Exploration and Visualization
  - Summary and stats
  - Various charts like pie charts and histograms
  - Exploration of multiple variables
  - Level plot, contour plot and 3D plot
  - Saving charts into files

Size and Structure of Data

- dim(iris)
- ## [1] 150 5
- names(iris)
- ## [1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Wid...
- ## [5] "Species"
- str(iris)
- ## 'data.frame': 150 obs. of 5 variables:
- ## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
- ## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1...
- ## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1...
- ## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0...
- ## $ Species : Factor w/ 3 levels "setosa","versicolor",....

Attributes of Data

- attributes(iris)

- ## $names
- ## [1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Wid...
- ## [5] "Species"
- ##
- ## $row.names
- ## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 ...
- ## [16] 16 17 18 19 20 21 22 23 24 25 26 27 28 ...
- ## [31] 31 32 33 34 35 36 37 38 39 40 41 42 43 ...
- ## [46] 46 47 48 49 50 51 52 53 54 55 56 57 58 ...
- ## [61] 61 62 63 64 65 66 67 68 69 70 71 72 73 ...
- ## [76] 76 77 78 79 80 81 82 83 84 85 86 87 88 ...
- ## [91] 91 92 93 94 95 96 97 98 99 100 101 102 103 1...
- ## [106] 106 107 108 109 110 111 112 113 114 115 116 117 118 1...
- ## [121] 121 122 123 124 125 126 127 128 129 130 131 132 133 1...
- ## [136] 136 137 138 139 140 141 142 143 144 145 146 147 148 1...
- ##
- ## $class
- ## [1] "data.frame"

First Rows of Data

- iris[1:3,]
- ## Sepal.Length Sepal.Width Petal.Length Petal.Width Species
- ## 1 5.1 3.5 1.4 0.2 setosa
- ## 2 4.9 3.0 1.4 0.2 setosa
- ## 3 4.7 3.2 1.3 0.2 setosa
- head(iris, 3)
- ## Sepal.Length Sepal.Width Petal.Length Petal.Width Species
- ## 1 5.1 3.5 1.4 0.2 setosa
- ## 2 4.9 3.0 1.4 0.2 setosa
- ## 3 4.7 3.2 1.3 0.2 setosa
- tail(iris, 3)
- ## Sepal.Length Sepal.Width Petal.Length Petal.Width Spe...
- ## 148 6.5 3.0 5.2 2.0 virgi...
- ## 149 6.2 3.4 5.4 2.3 virgi...
- ## 150 5.9 3.0 5.1 1.8 virgi...

A Single Column

- The first 10 values of Sepal.Length
- iris[1:10, "Sepal.Length"]
- ## [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9
- iris$Sepal.Length[1:10]
- ## [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9

Summary of Data

- Function summary()
  - numeric variables: minimum, maximum, mean, median, and the first (25%) and third (75%) quartiles
  - categorical variables (factors): frequency of every level
- summary(iris)
- ## Sepal.Length Sepal.Width Petal.Length Petal.Wid...
- ## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0....
- ## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0....
- ## Median :5.800 Median :3.000 Median :4.350 Median :1....
- ## Mean :5.843 Mean :3.057 Mean :3.758 Mean :1....
- ## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1....
- ## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2....
- ## Species
- ## setosa :50
- ## versicolor:50
- ## virginica :50

Mean, Median, Range and Quartiles

- Mean, median and range: mean(), median(), range()
- Quartiles and percentiles: quantile()

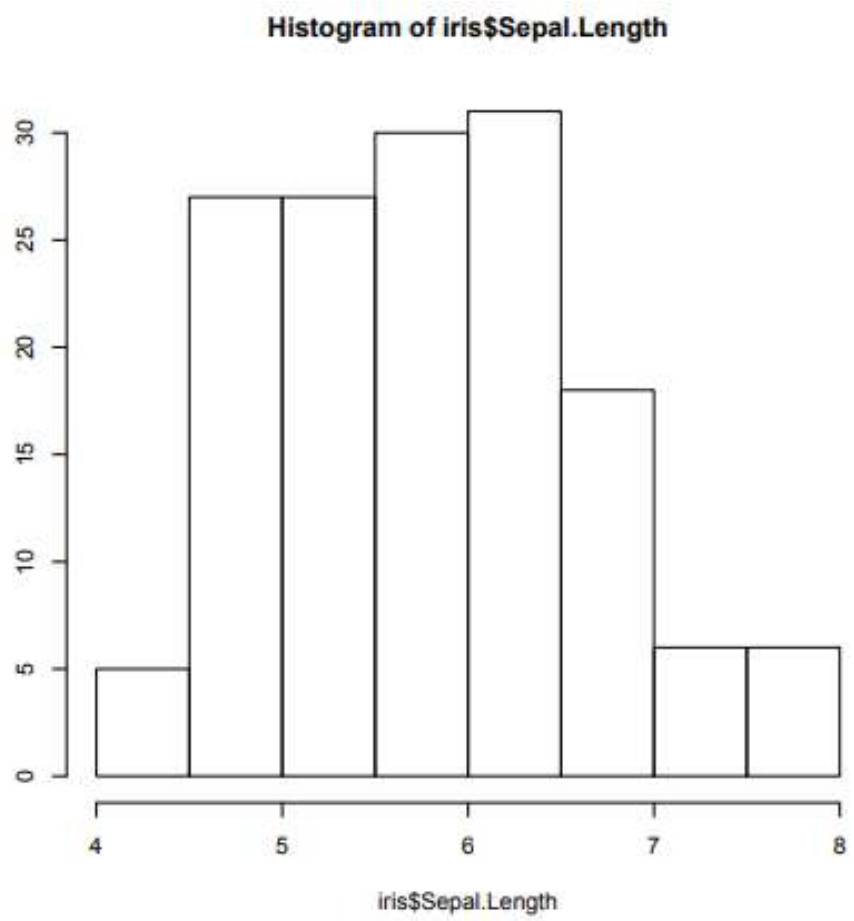
- range(iris$Sepal.Length)
- ## [1] 4.3 7.9

- quantile(iris$Sepal.Length)
- ## 0% 25% 50% 75% 100%
- ## 4.3 5.1 5.8 6.4 7.9

- quantile(iris$Sepal.Length, c(0.1, 0.3, 0.65))
- ## 10% 30% 65%
- ## 4.80 5.27 6.20

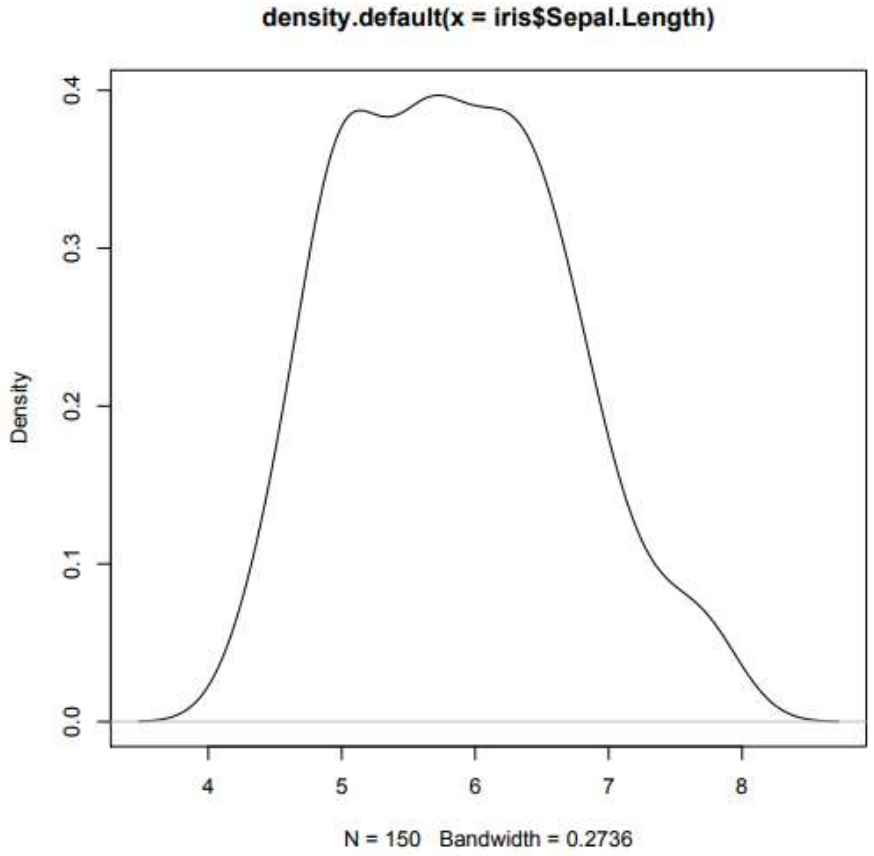
Variance and Histogram

- var(iris$Sepal.Length)
- ## [1] 0.6856935
- hist(iris$Sepal.Length)



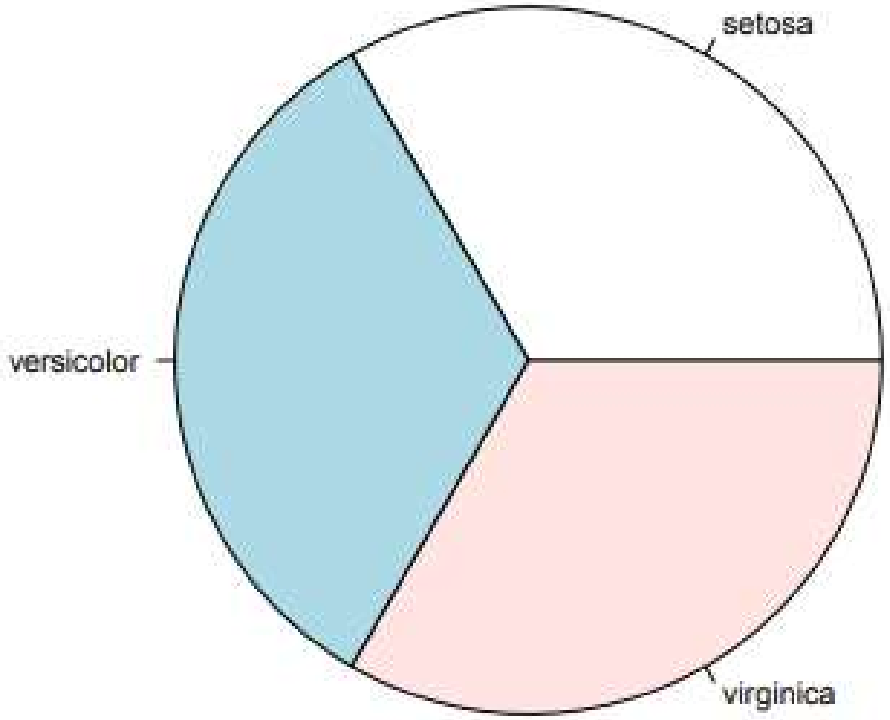
Density

- plot(density(iris$Sepal.Length))



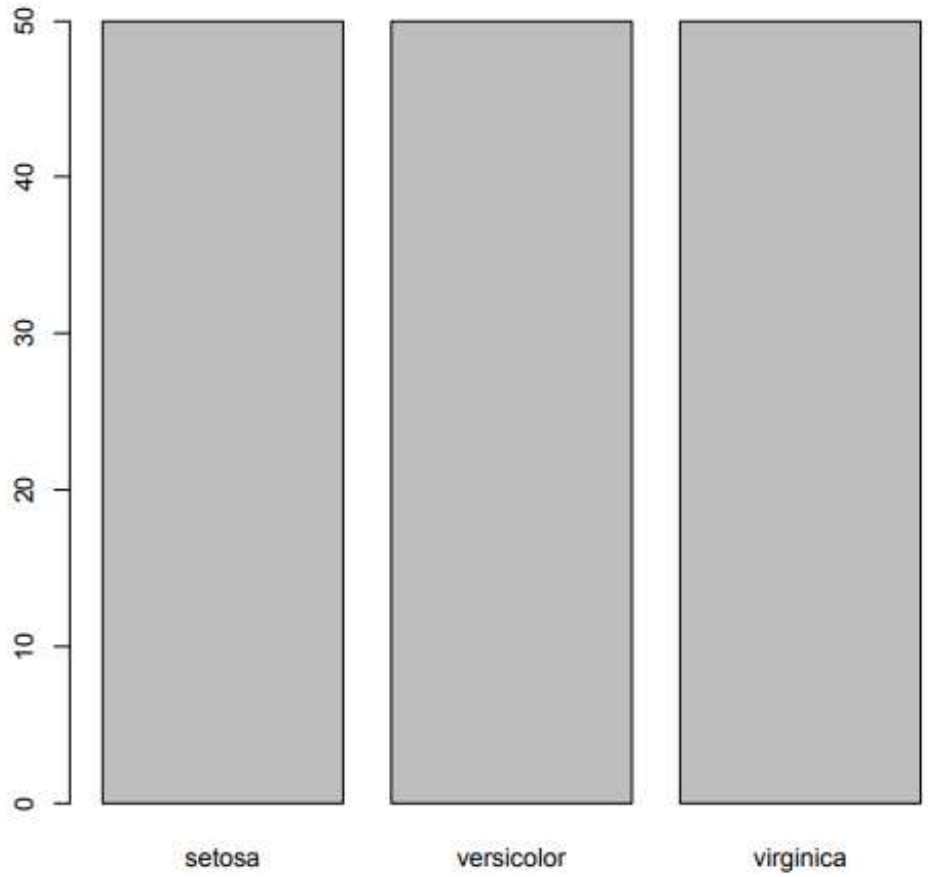
Pie Chart

- Frequency of factors: table()
- table(iris$Species)
  - ##
  - ## setosa versicolor virginica
  - ## 50 50 50
- pie(table(iris$Species))



Bar Chart

- barplot(table(iris$Species))




Thank
you !!!