UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

HDAT9800

Visualisation and Communication of Health Data

Chapter 3 - Basic data visualisation in R, with emphasis on ggplot2 - part A

Drs James Farrow & Tim Churches

Outline of Chapter 3

- pragmatic focus on how to create data visualisations in R
- with a little bit of underlying theory
- quick overview of the base R graphics and lattice graphics systems
- then focus on ggplot2
  - in subsequent chapters we'll look at add-ons and extensions to the *ggplot2* framework
  - as well as a bit more theory on good visualisation practice

ggplot2

The *ggplot2* package for R (part of the tidyverse) facilitates the creation of sophisticated layered statistical graphics

It uses an underlying *grammar* of graphics which allows graphs to be built up component-by-component

This approach allows individual aspects of a graph, such as the colour of one particular data series or the transparency of an overlay, to be easily varied independently of other aspects of a graph.

Base R graphics

R comes with extensive built-in data visualisation capabilities

- as you would expect from a data analysis system which focuses on statistical analysis

It dates back to S and S-Plus, the predecessor to R from the 1980s

- the outputs are publication-quality, with a nod to Tufte principles
- the graphics functions leverage the object-oriented nature of R
  - a given function, say *plot()*, behaves differently depending on the class of the object you pass to it
  - but the graphics themselves are not first-class objects that can be further manipulated in R

Base graphics in R

Don't ignore base graphics function in R!

- still fit-for-purpose, especially exploratory data analysis (EDA)
- many high-level features for diagnostics of statistical models and special-purpose visualisations
- ideal for a quick, interactive use at the R console command line
- easy-to-understand "*artist's palette*" model of plotting:
    - you start with a blank canvas and build up your chart from there

Base graphics in R

Resources:

- *An Introduction to R Graphics* - a detailed history and lots of amazing examples of base R graphics by Paul Murrell at University of Auckland.

- *Chapter 6* and *Section 7.1* in the excellent online book *Exploratory Data Analysis in R* by Roger Peng

- *R Base Graphics: An Idiot's Guide* by Susan Johnstone is also a very useful introduction (code and data files for it available here)

The slides which follow make us of examples from these resources.

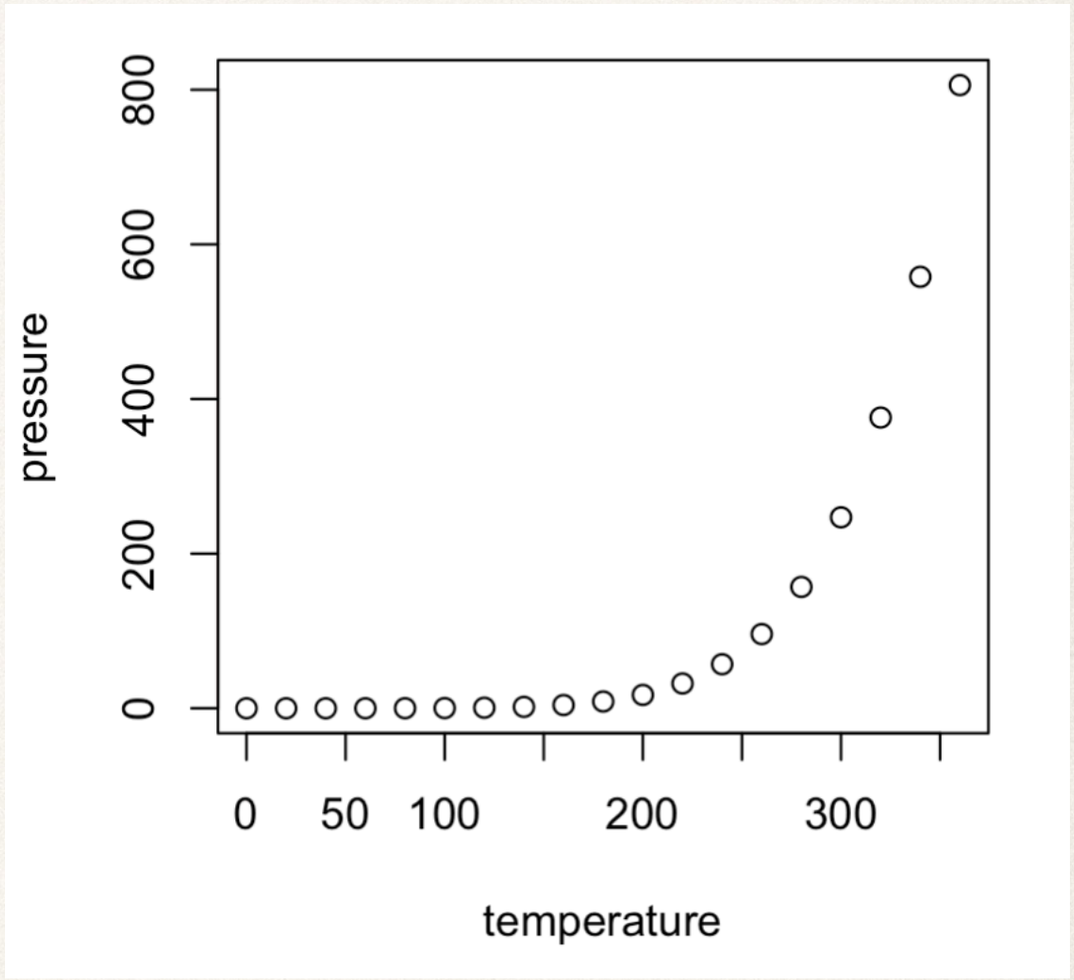
Base graphics in R: `plot()`

Built-in dataset of the vapour pressure of mercury

```
> pressure
   temperature pressure
1            0   0.0002
2           20   0.0012
3           40   0.0060
4           60   0.0300
5           80   0.0900
6          100   0.2700
7          120   0.7500
8          140   1.8500
9          160   4.2000
10         180   8.8000
11         200  17.3000
12         220  32.1000
13         240  57.0000
14         260  96.0000
15         280 157.0000
16         300 247.0000
17         320 376.0000
18         340 558.0000
19         360 806.0000
>
```

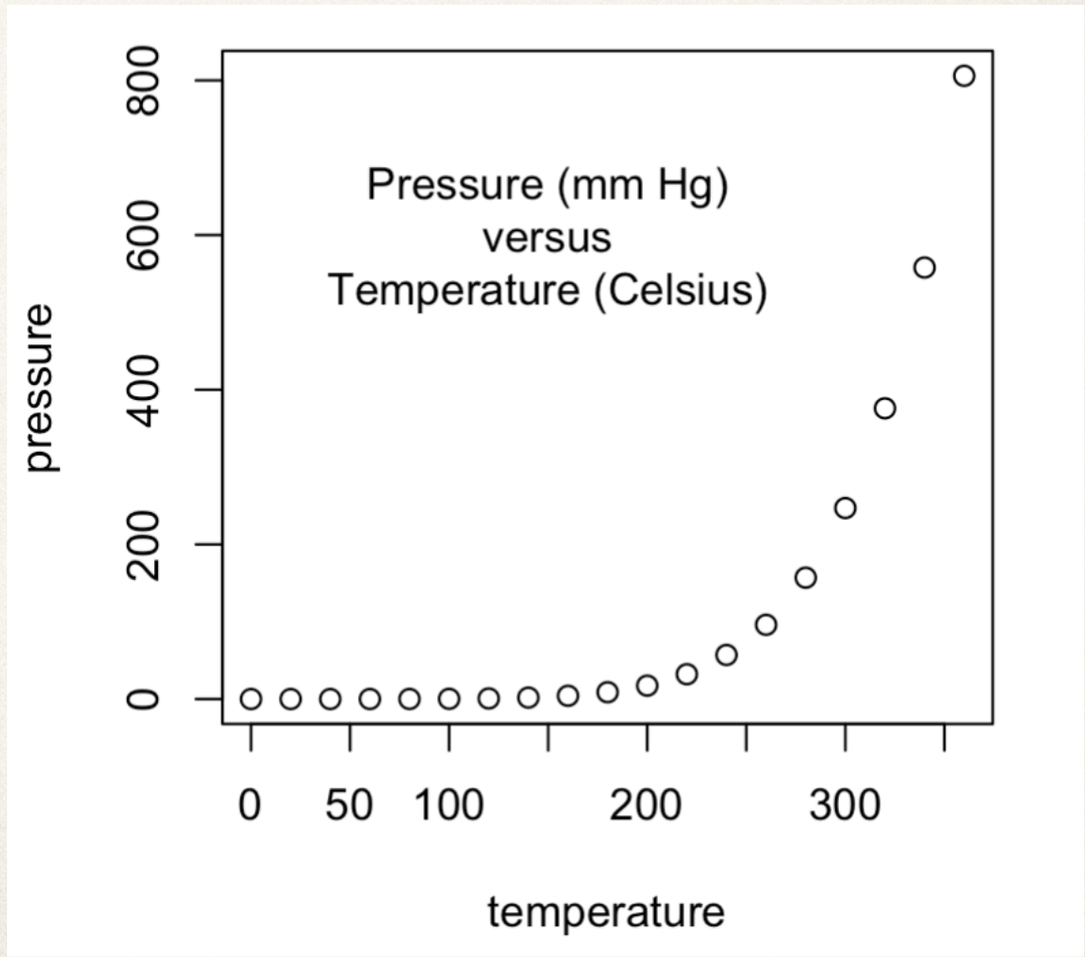
Base graphics in R: `plot()`

```
> plot(pressure)
```



Base graphics in R: `plot()`

```
> plot(pressure)

> text(150, 600,”Pressure (mm Hg)\nversus\nTemperature
(Celsius)")
```

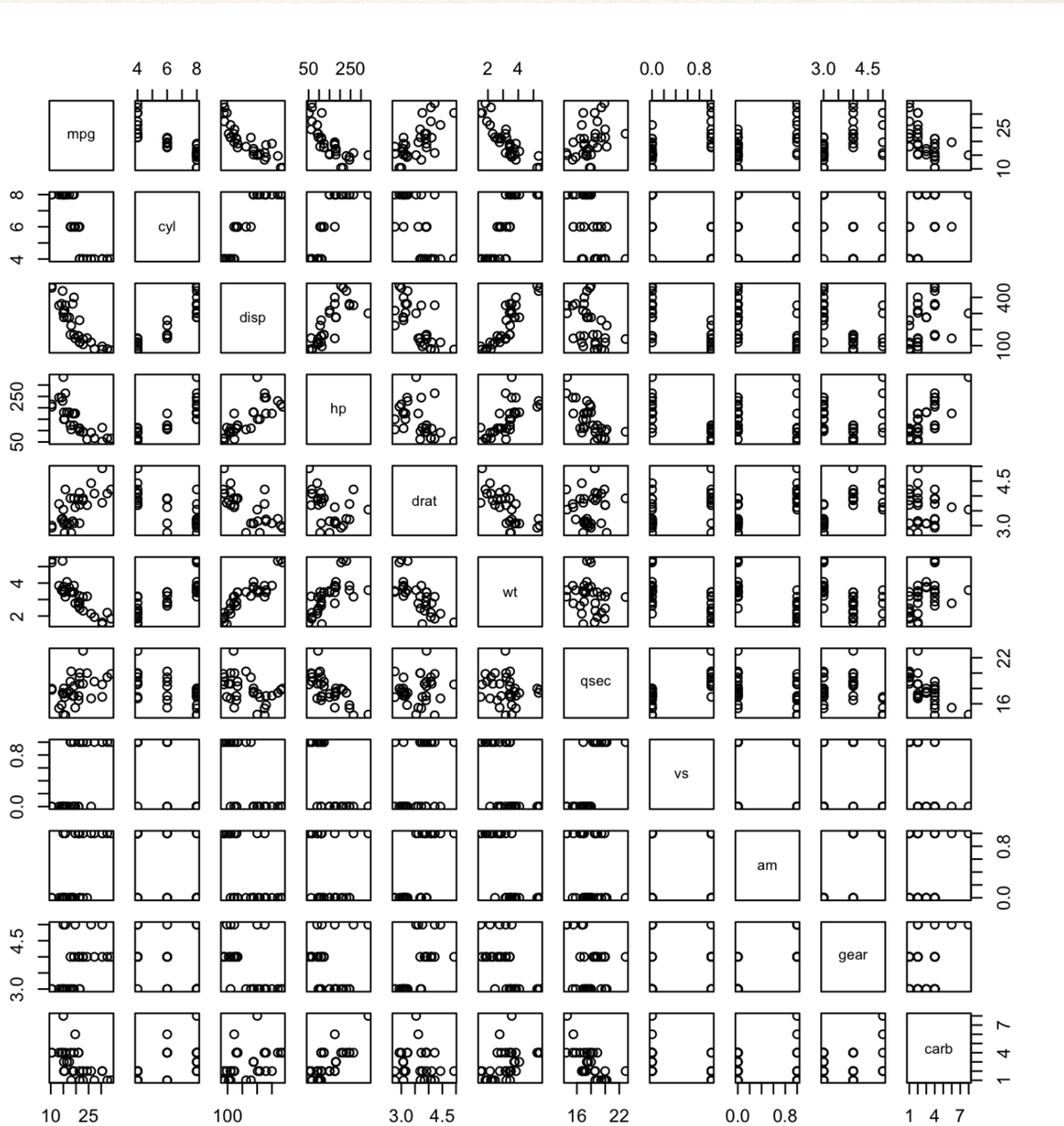


Base graphics in R: `plot()`

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

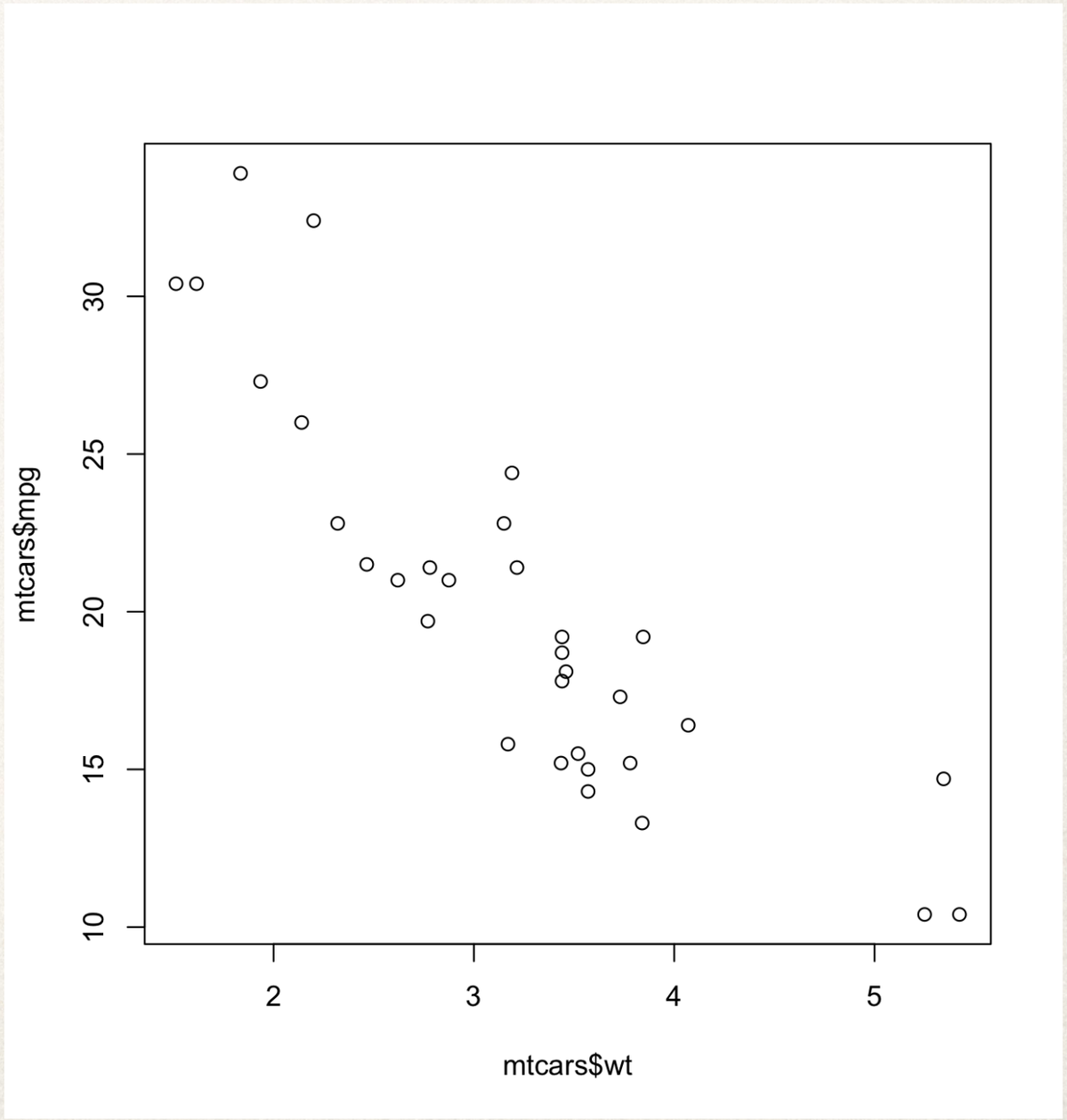
Base graphics in R: `plot()`

```
> plot(mtcars)
```



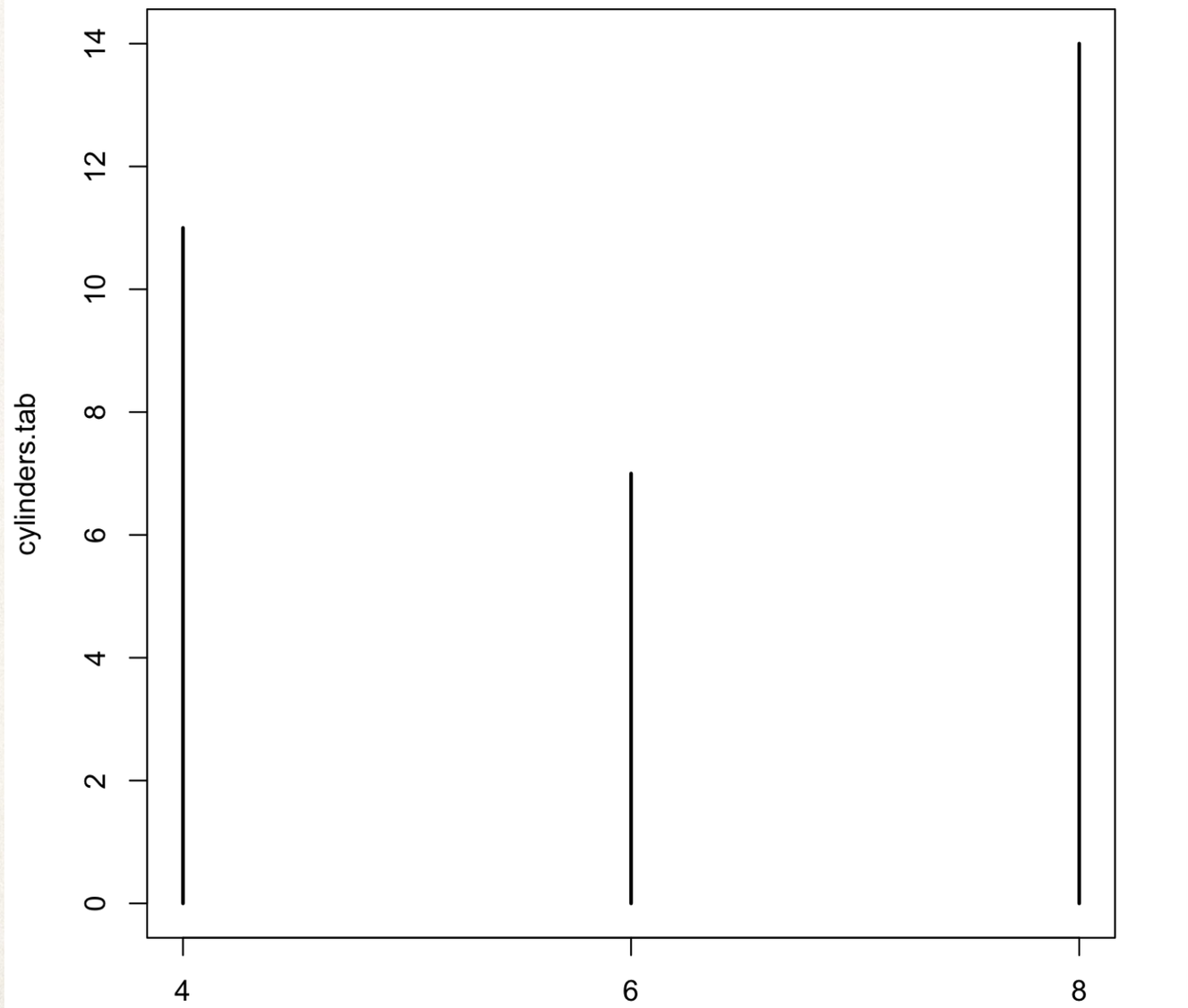
Base graphics in R: `plot()`

```
> plot(mtcars$wt, mtcars$mpg)
```



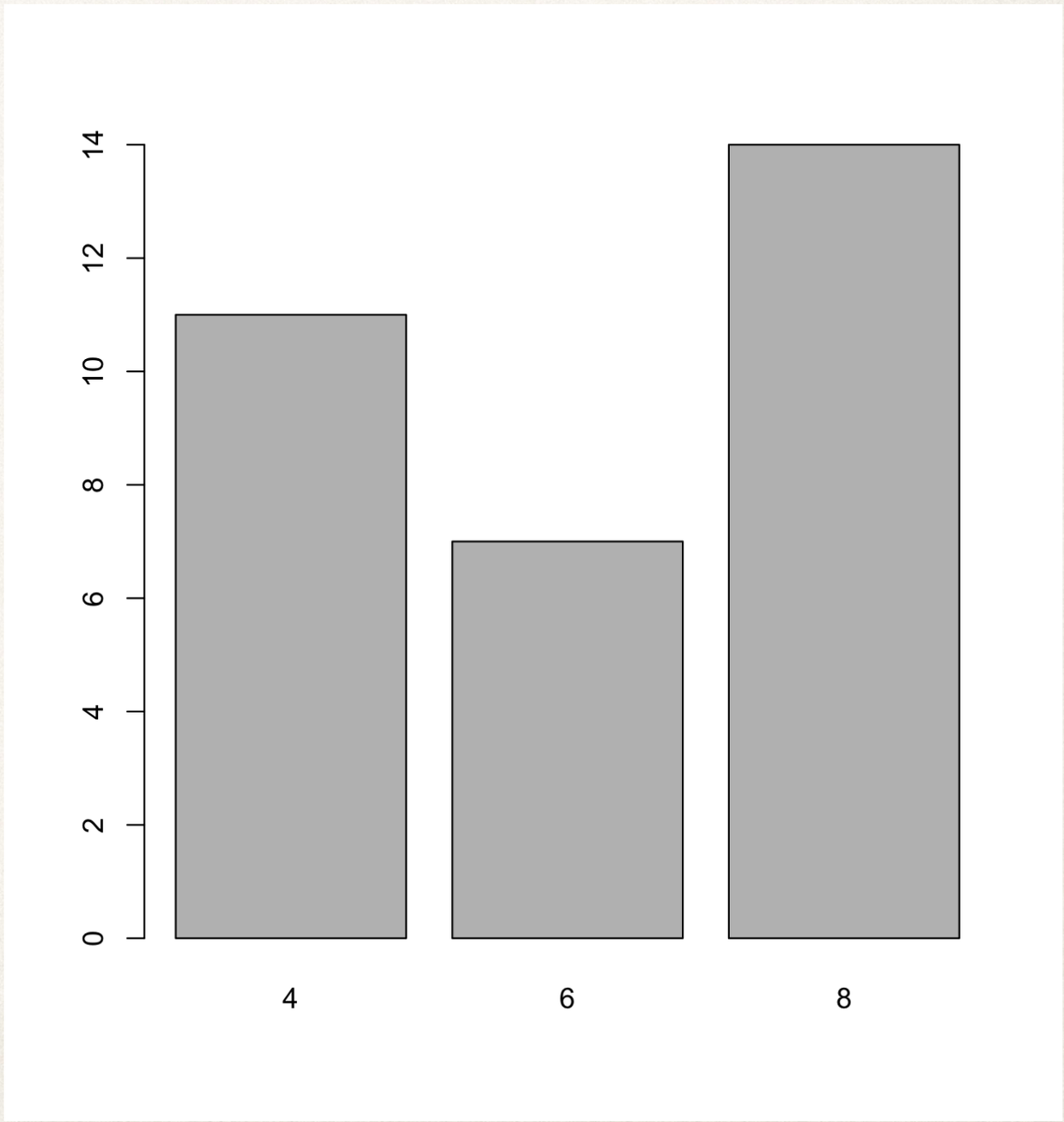
Base graphics in R: `plot()`

```
> cylinders.tab <- table(mtcars$cyl)

> cylinders.tab

 4  6  8

11  7 14

> plot(cylinders.tab)
```



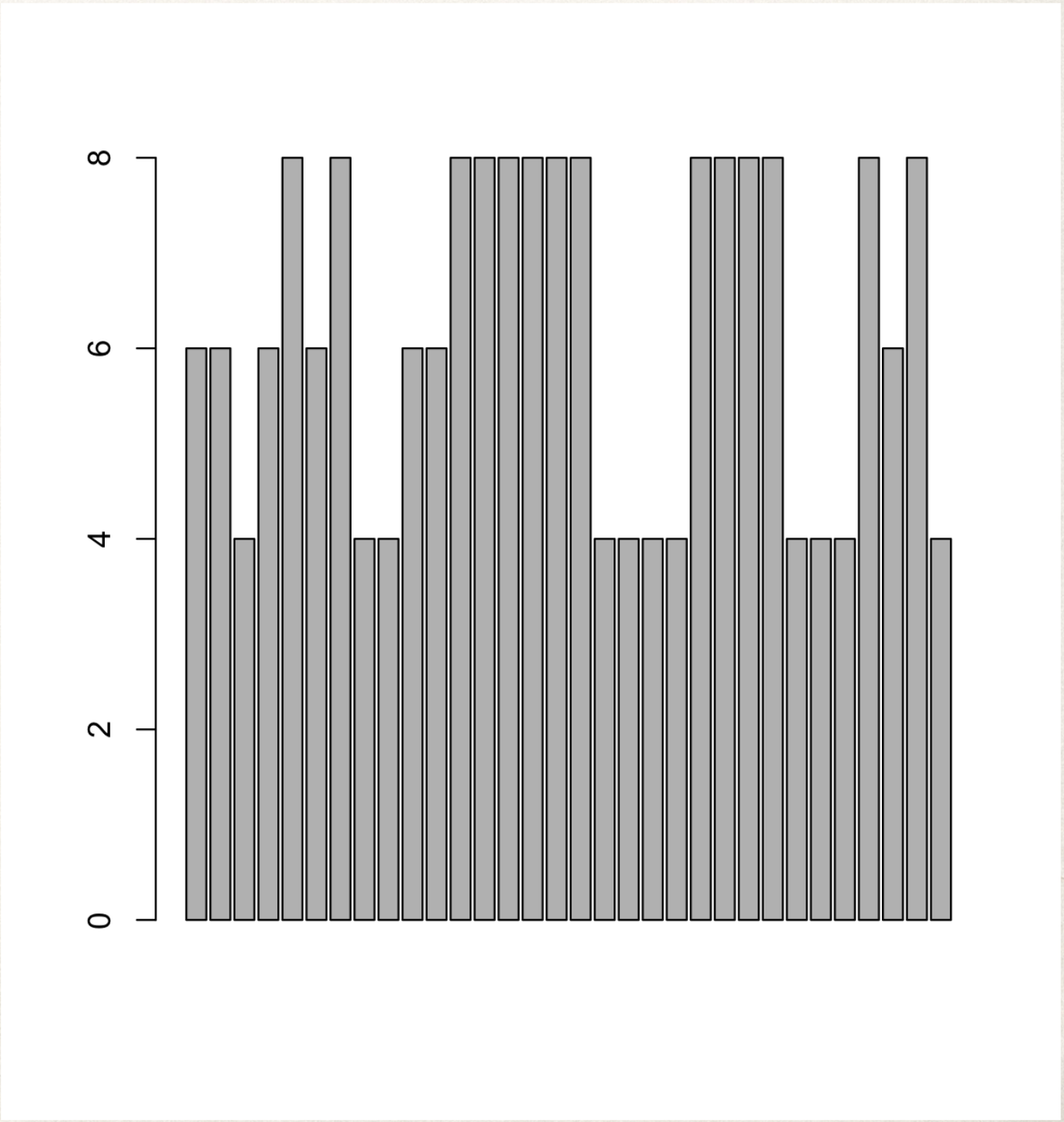
Base graphics in R: `barplot()`

```
> cylinders.tab <- table(mtcars$cyl)

> cylinders.tab

 4  6  8

11  7 14

> barplot(cylinders.tab)
```



Base graphics in R: `barplot()`

```
> barplot(mtcars$cyl)
```

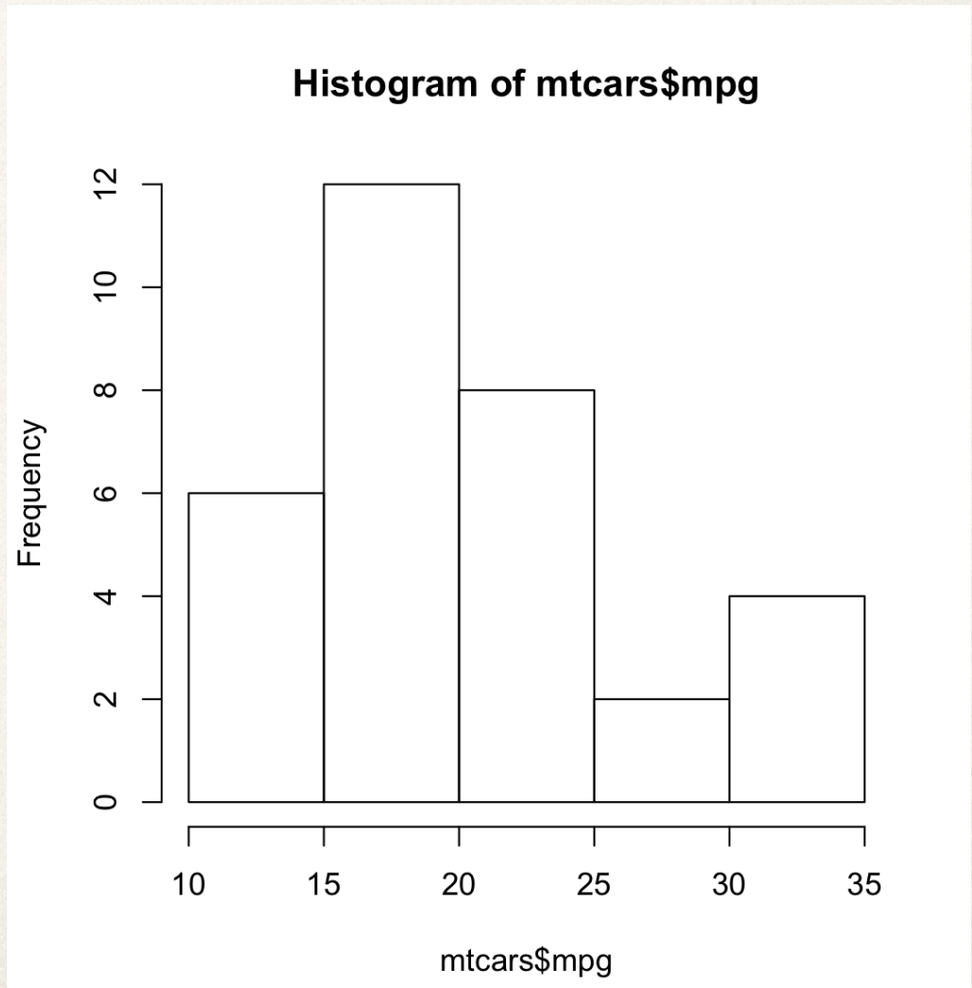


Base graphics in R: `hist()`

```
> hist(mtcars)
```

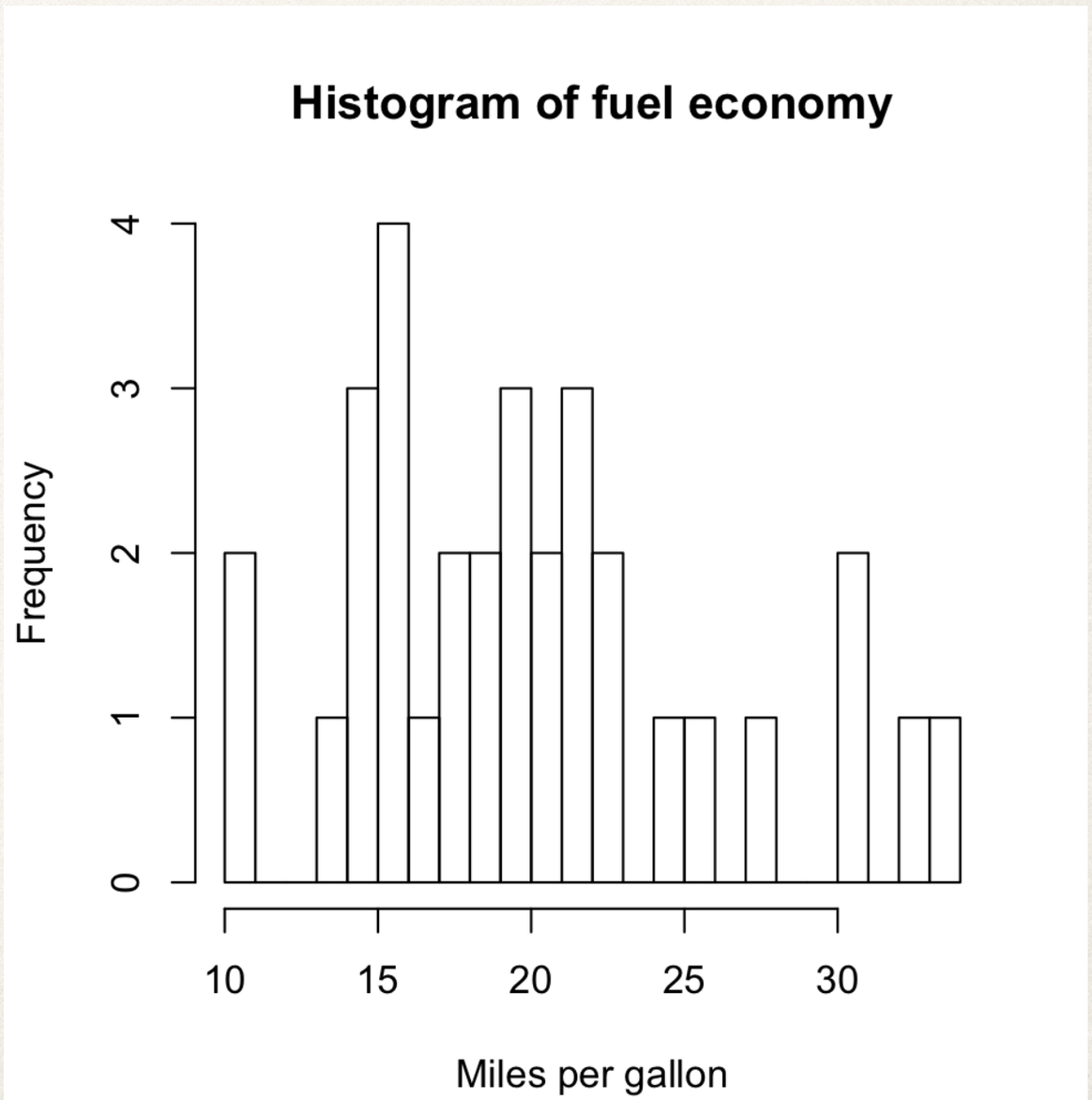
```
Error in hist.default(mtcars) : 'x' must be numeric
```

```
> hist(mtcars$mpg)
```



Base graphics in R: `hist()`

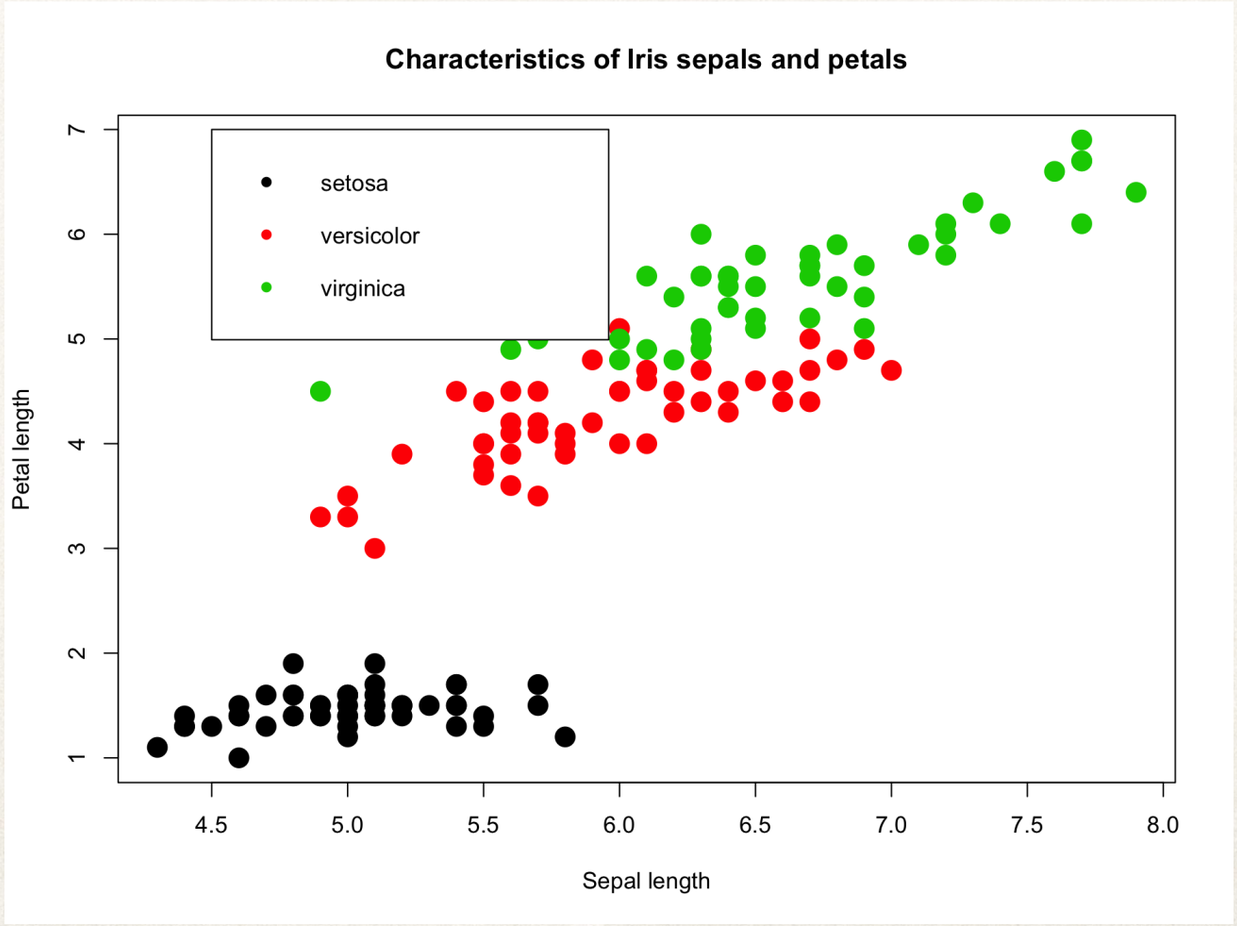
```
> hist(mtcars$mpg, breaks=20, xlab="Miles per gallon", main="Histogram of fuel
economy")
```



Base graphics in R: scatter plots

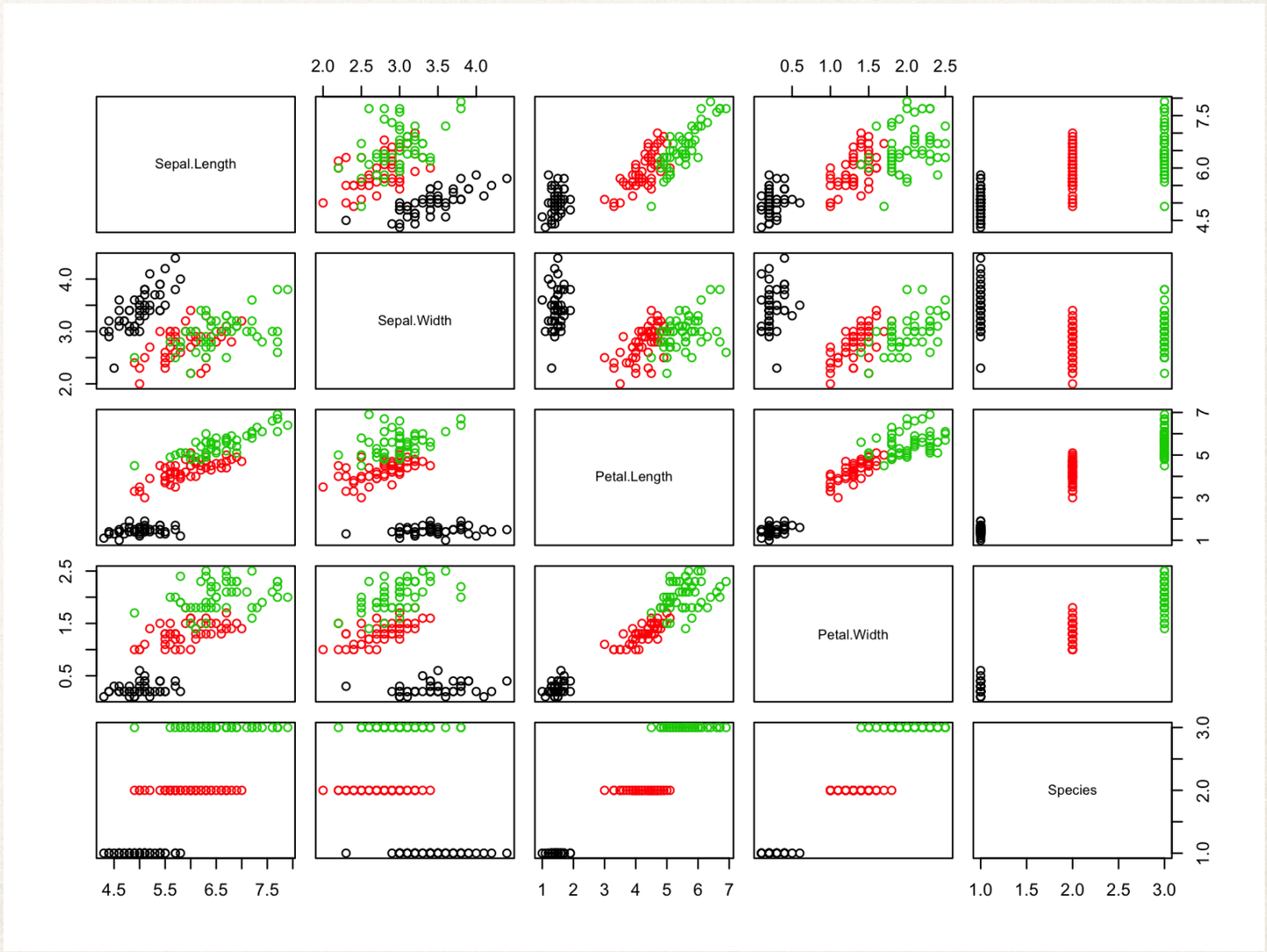
```
> plot(iris$Sepal.Length, iris$Petal.Length, col=iris$Species, pch=16, cex=2,
xlab="Sepal length", ylab="Petal length", main="Characteristics of Iris sepals
and petals")
```

```
> legend(x=4.5, y=7, legend=levels(iris$Species), col=1:3, pch=16)
```



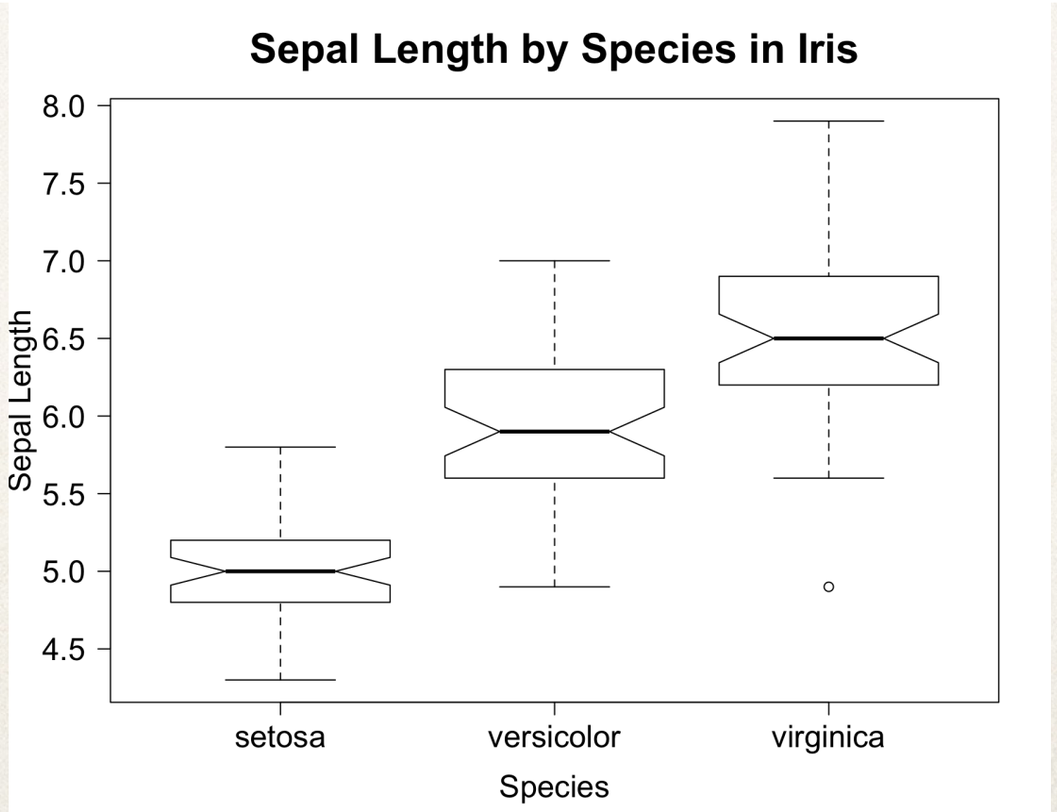
Base graphics in R: scatter plots

```
pairs(iris, col=iris$Species)
```

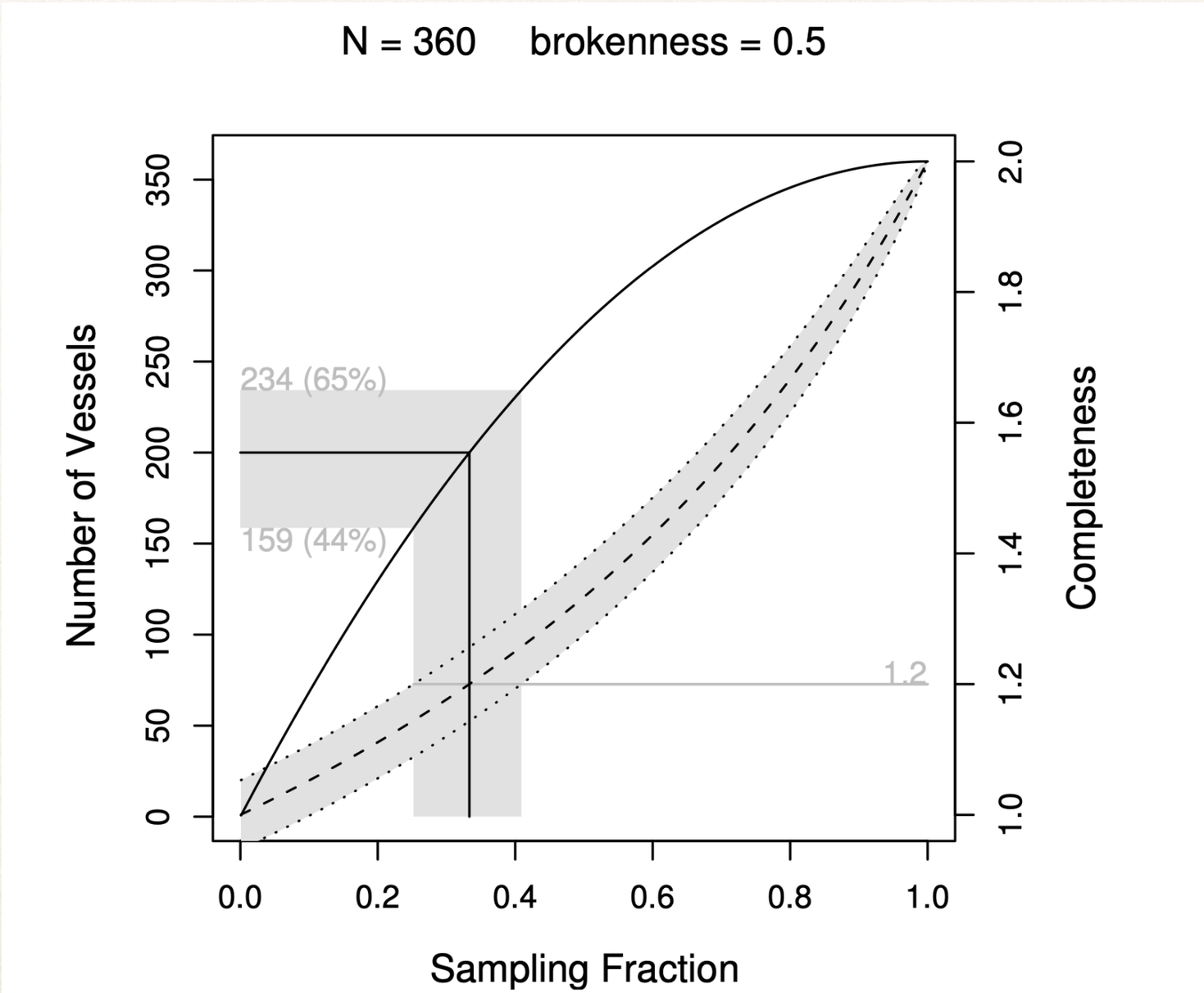


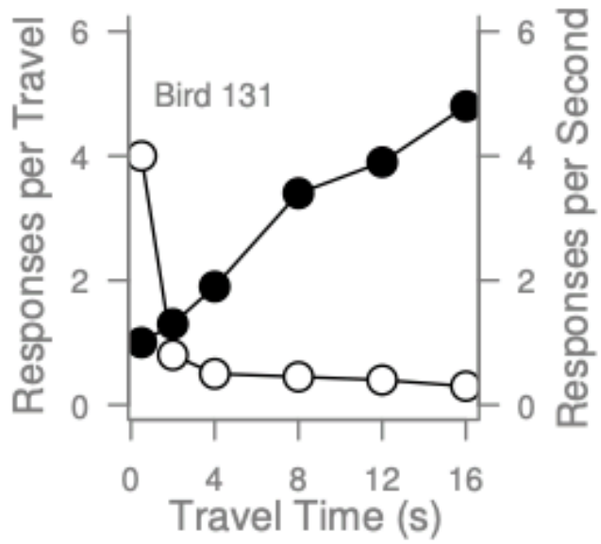
Base graphics in R: `boxplot()`

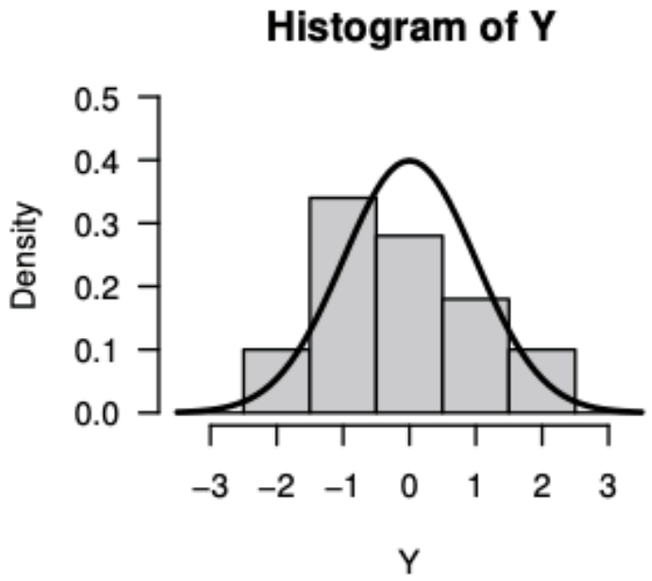
```
> boxplot(iris$Sepal.Length ~ iris$Species,                  # x variable, y variable
          notch = T,                                         # Draw notch
          las = 1,                                           # Orientate the axis tick labels
          xlab = "Species",                                  # X-axis label
          ylab = "Sepal Length",                             # Y-axis label
          main = "Sepal Length by Species in Iris",          # Plot title
          cex.lab = 1.5,                                     # Size of axis labels
          cex.axis = 1.5,                                    # Size of the tick mark labels
          cex.main = 2)                                      # Size of the plot title
```

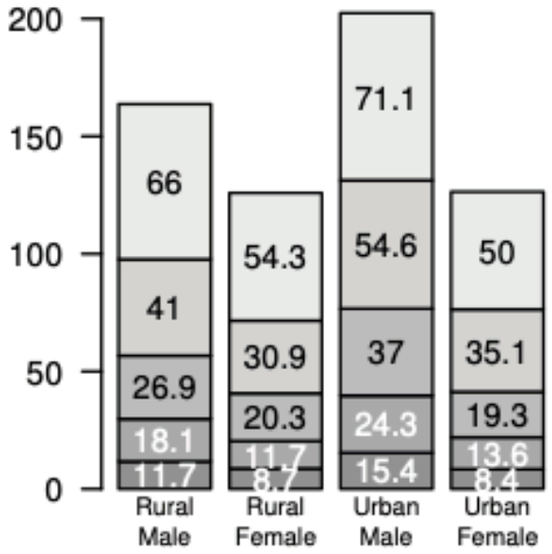


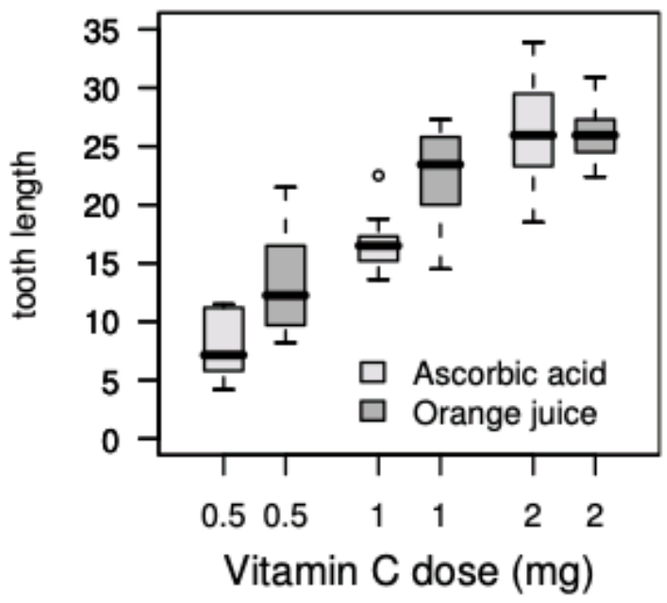
Base graphics in R: `overlays`

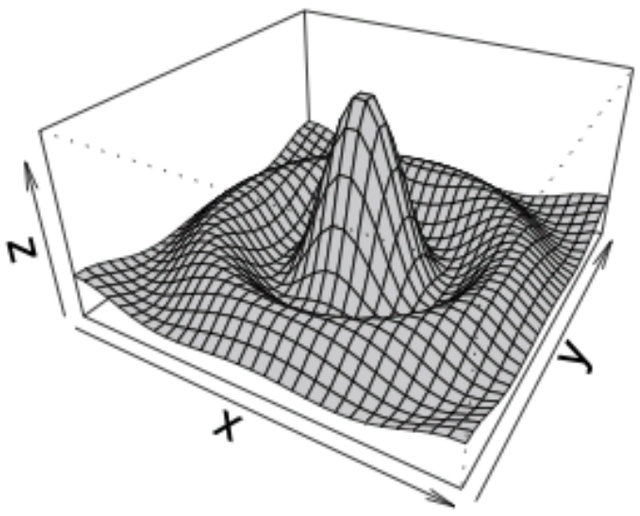


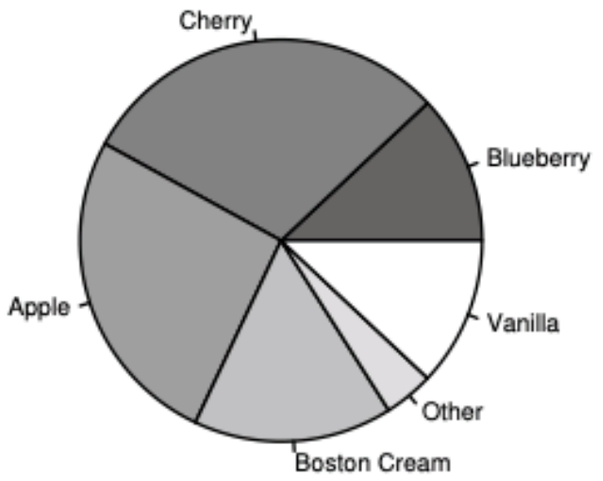

Bird 131
Responses per Travel
Responses per Second
Travel Time (s)
0
2
4
6
0
4
8
12
16

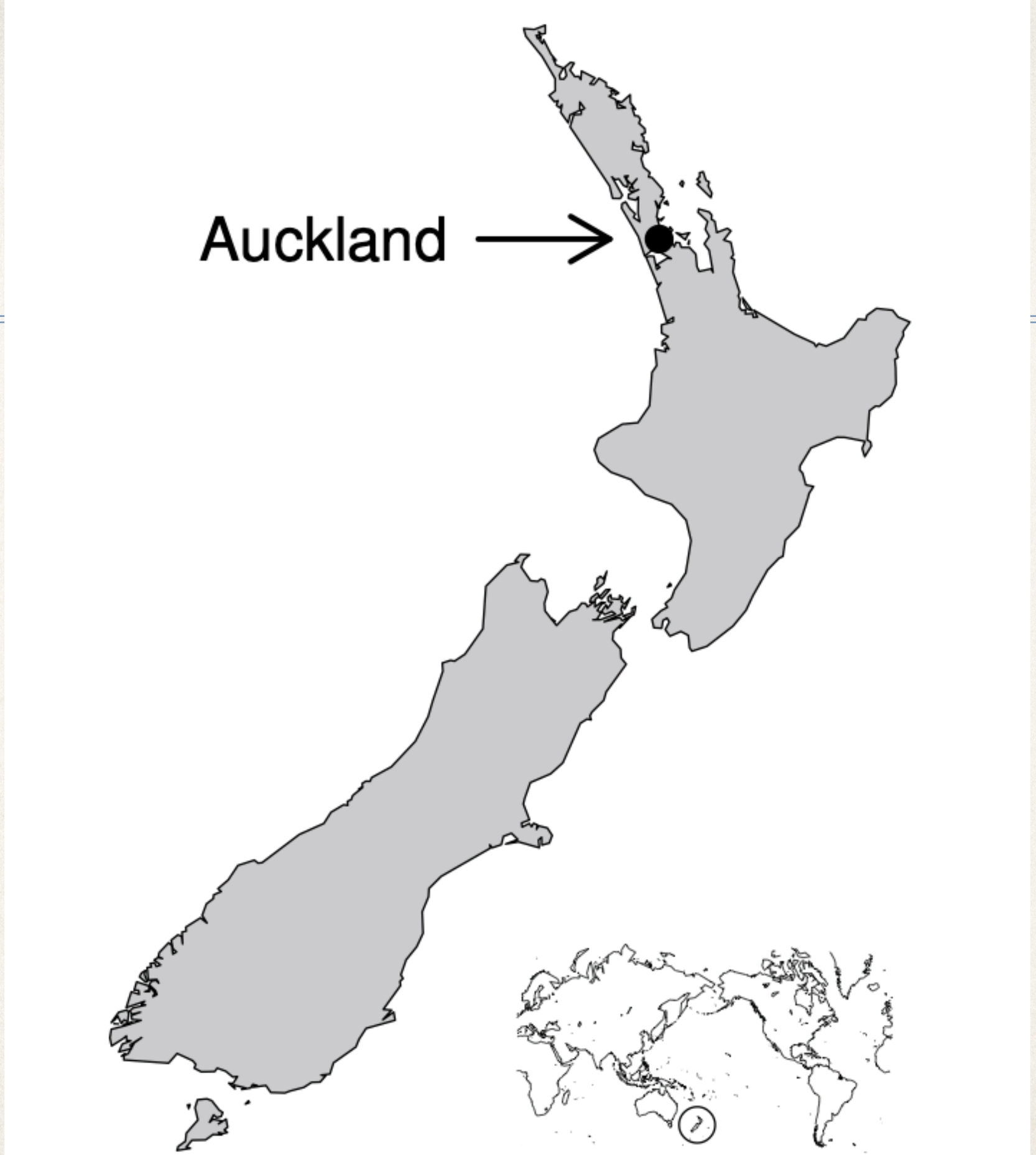

Histogram of Y
Density
Y
0.0
0.1
0.2
0.3
0.4
0.5
−3
−2
−1
0
1
2
3

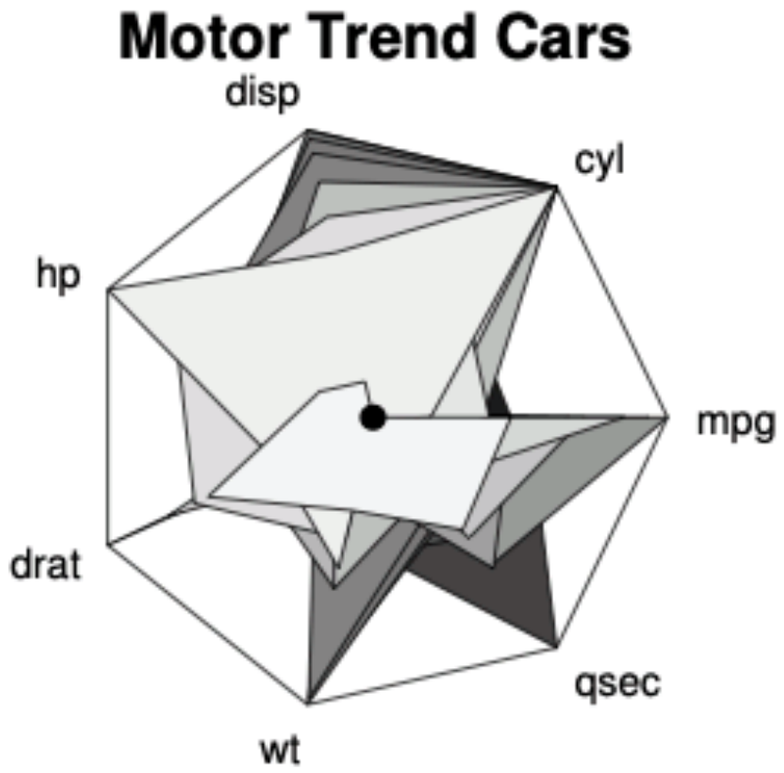

66
41
26.9
18.1
11.7
54.3
30.9
20.3
11.7
8.7
71.1
54.6
37
24.3
15.4
50
35.1
19.3
13.6
8.4
0
50
100
150
200
Rural Male
Rural Female
Urban Male
Urban Female

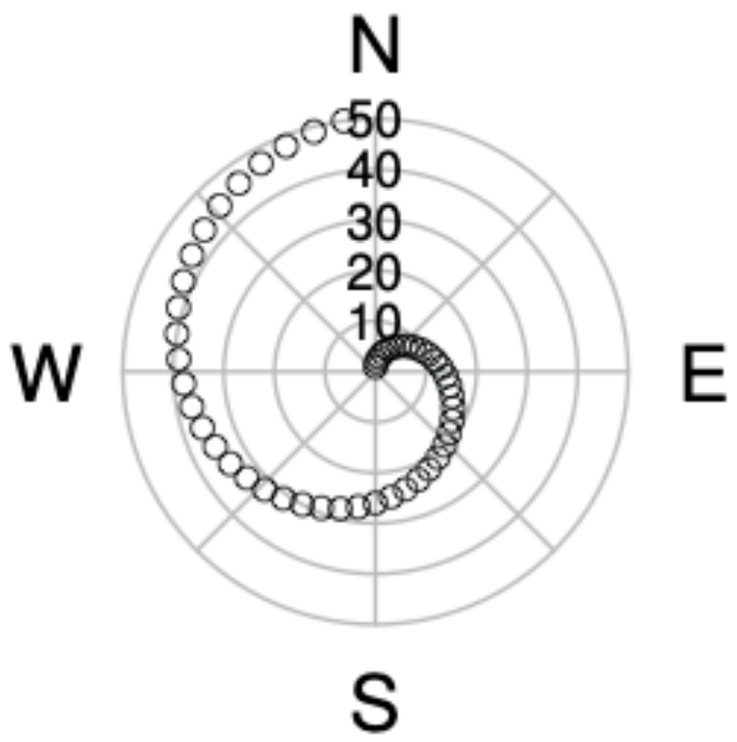

tooth length
Vitamin C dose (mg)
Ascorbic acid
Orange juice
0
5
10
15
20
25
30
35
0.5
0.5
1
1
2
2

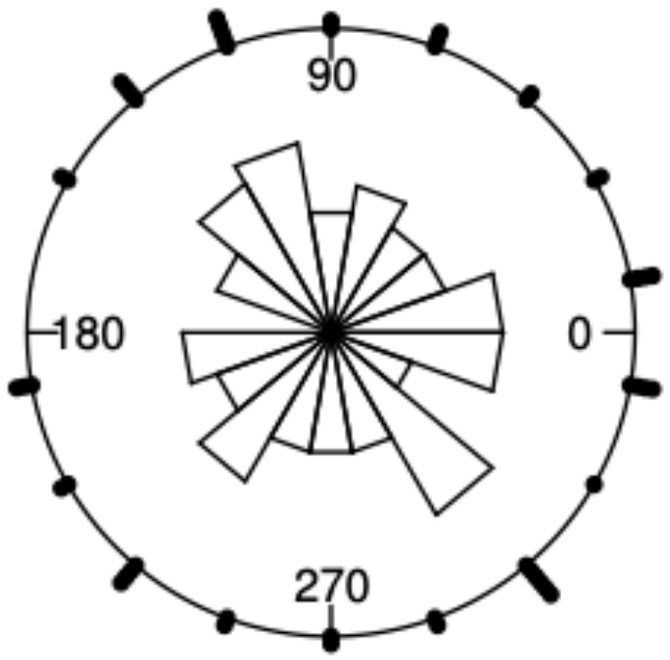

z
x
y

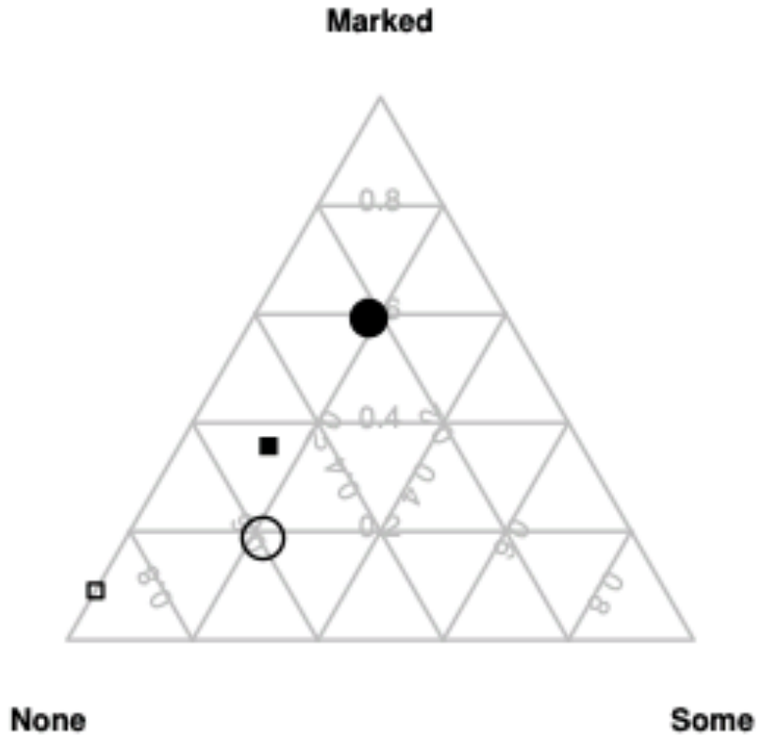

Cherry
Blueberry
Vanilla
Other
Boston Cream
Apple

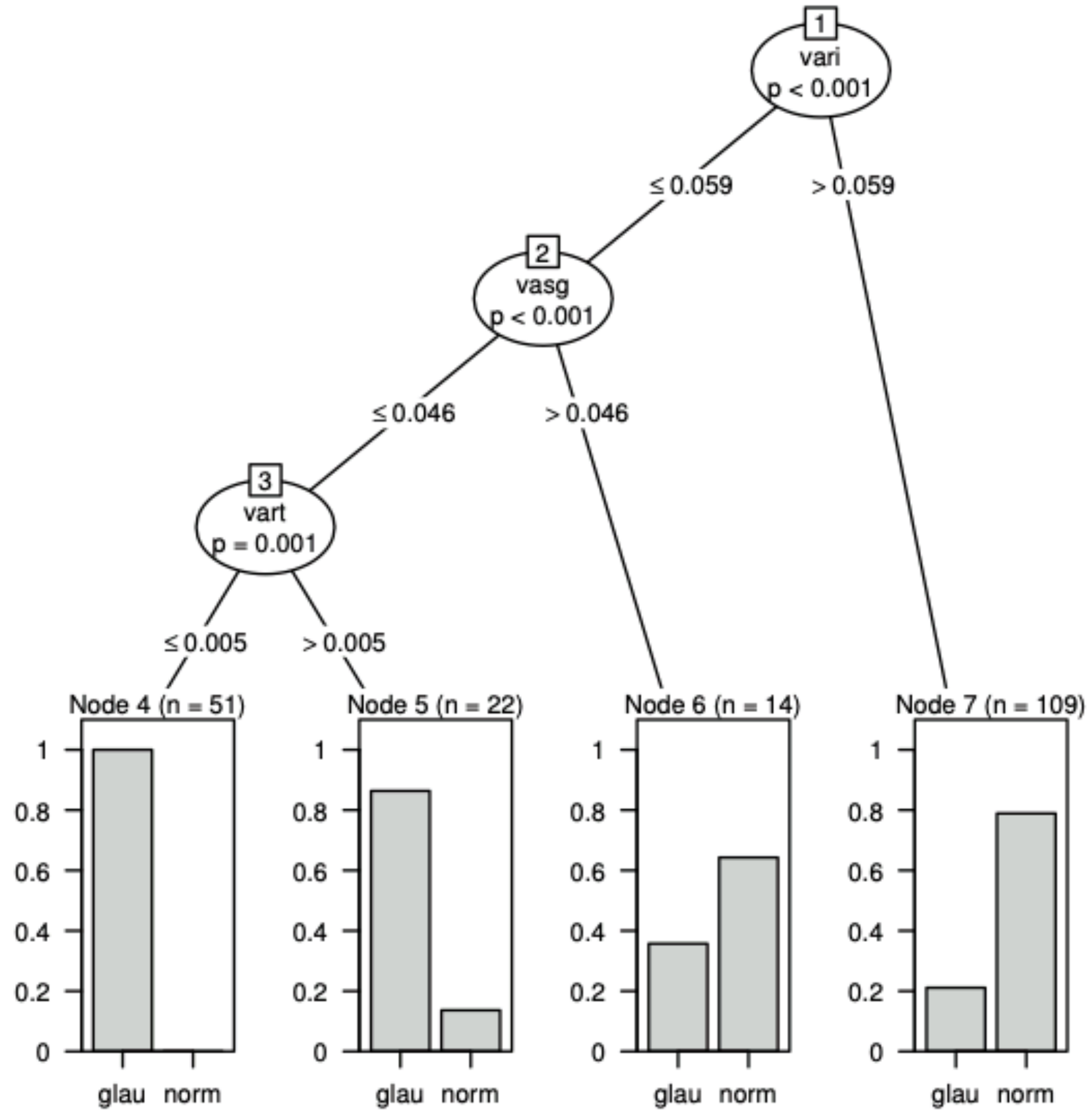

Auckland

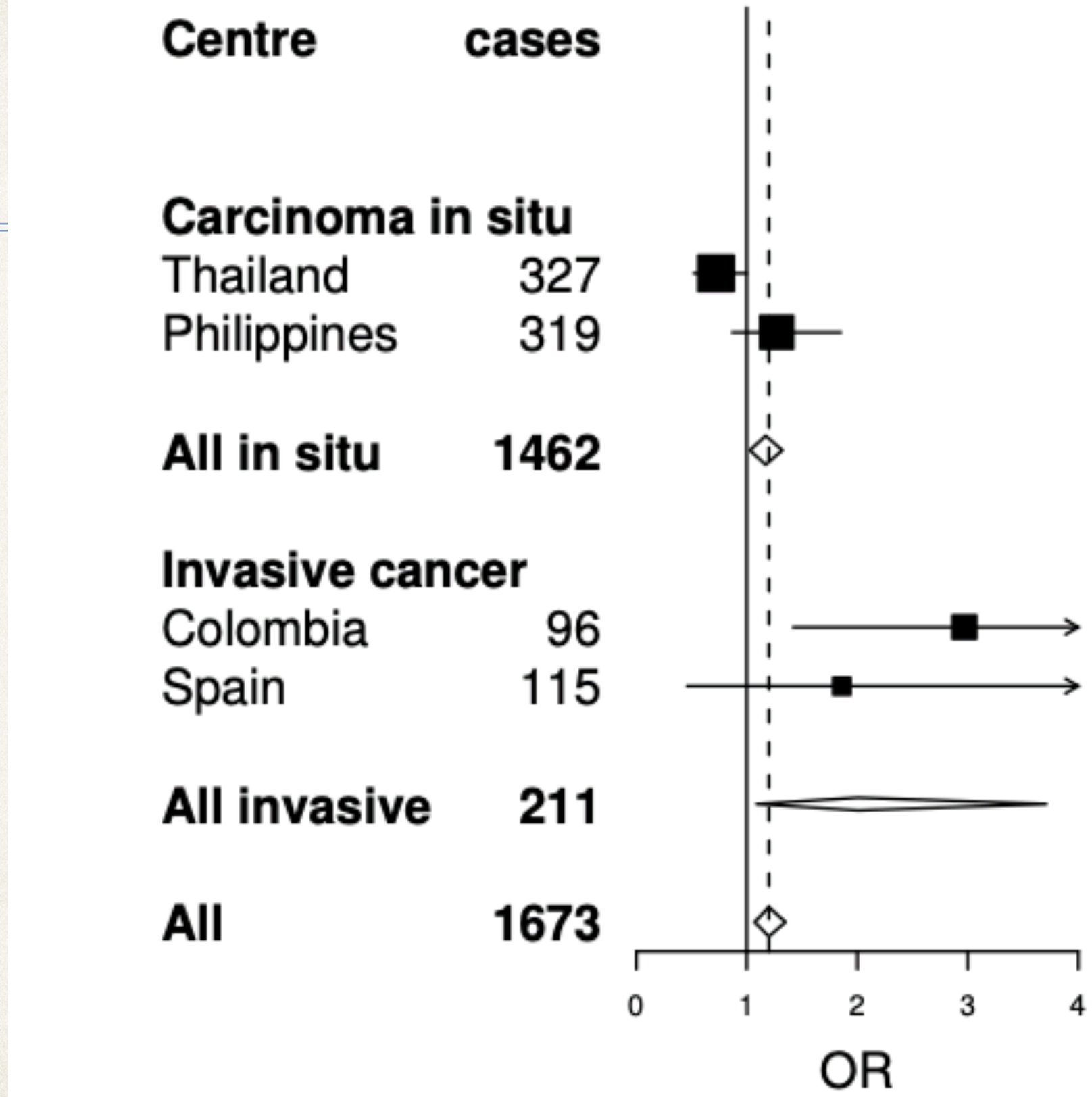

Motor Trend Cars
disp
cyl
hp
mpg
drat
qsec
wt


N
50
40
30
20
10
W
E
S


90
180
0
270


Marked
0.8
0.4
None
Some


1
vari
p < 0.001
≤ 0.059
> 0.059
2
vasg
p < 0.001
≤ 0.046
> 0.046
3
vart
p = 0.001
≤ 0.005
> 0.005
Node 4 (n = 51)
Node 5 (n = 22)
Node 6 (n = 14)
Node 7 (n = 109)
1
0.8
0.6
0.4
0.2
0
glau
norm


Centre
cases
Carcinoma in situ
Thailand 327
Philippines 319
All in situ 1462
Invasive cancer
Colombia 96
Spain 115
All invasive 211
All 1673
0
1
2
3
4
OR

Base graphics in R: a summary

With enough effort, almost any type of data visualisation can be created using base graphics in R, but…

- some common actions are not straightforward (legends, faceting, error bars
- complex graphs are time-consuming
- the resulting plots can't be further manipulated by R code

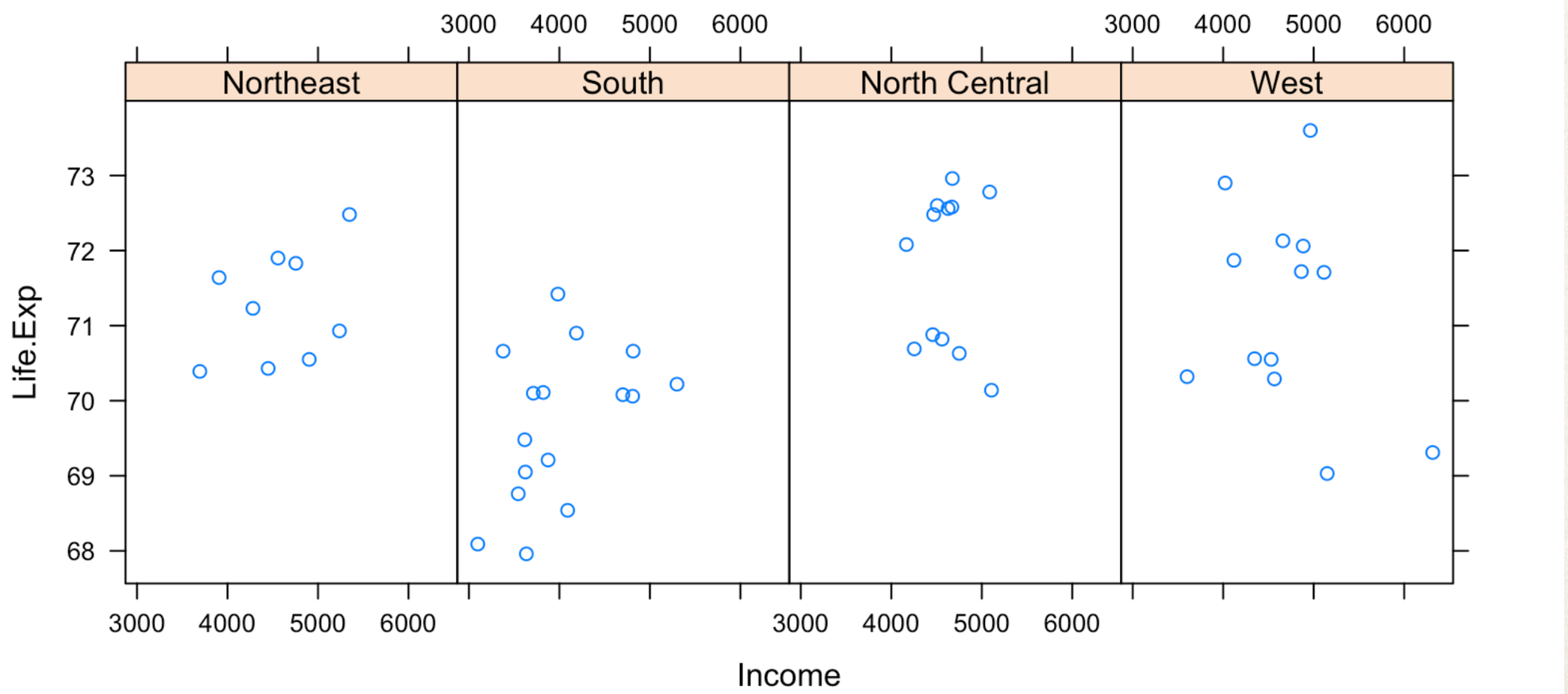
Therefore, base graphics are great for quick exploratory data analysis, and for some special purposes, including diagnostics for statistical models, but…

- ggplot2 is superior for almost everything else

Lattice graphics in R

`lattice` is another graphics package for R which became popular in the mid-2000s, (prior to `ggplot2`)

- implemented some key ideas proposed by Bill Clevland, particularly trellis plots
- uses a formula interface, same as many R modelling functions
- but is now largely superseded: everything you can do with the `lattice` package you can do better with ggplot2
- the resulting plots can't be further manipulated by R code

Lattice graphics in R

```
> state <- data.frame(state.x77, region = state.region)
> xyplot(Life.Exp ~ Income | region, data = state, layout = c(4, 1))
```