Data Analysis & Graphics Using R – Solutions to Selected Exercises (March 21, 2004)

Preliminaries

```
> library(DAAG)
```

Exercise 2
The following table gives the size of the floor area (ha) and the price ($000), for 15 houses sold in the Canberra (Australia) suburb of Aranda in 1999.
. . . . . .
Type these data into a data frame with column names `area` and `sale.price`.

(a) Plot `sale.price` versus `area`.

(b) Use the `hist()` command to plot a histogram of the sale prices.

(c) Repeat (a) and (b) after taking logarithms of sale prices.

The Aranda house price data are also in a data frame in the `DAAG` package, called `houseprices`.

(a) Omitted

(b) Omitted

(c) The following code demonstrates the use of the `log="y"` argument to cause `plot` to use a logarithmic scale on the `y` axis, but with axis tick labels that are specified in the original units.

```
> plot(sale.price ~ area, data = houseprices, log = "y", pch = 16,
+     xlab = "Floor Area", ylab = "Sale Price", main = "(c) log(sale.price) vs area")
```

The following puts a logarithmic scale on the x-axis of the histogram.

```
> hist(log(houseprices$sale.price), xlab = "Sale Price (logarithmic scale)",
+     main = "(d)  Histogram of log(sale.price)")
```

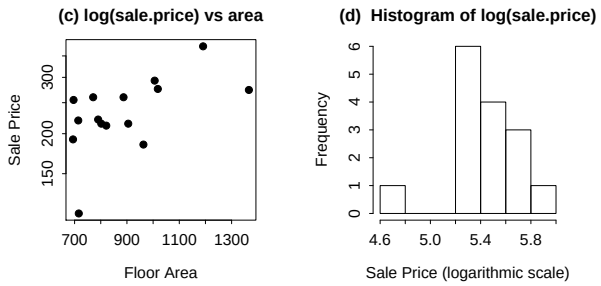

Figure 1: Plots for Exercise 2c.

Here is an alternative that prints x-axis labels in the original units:

```
> logbreaks <- hist(log(houseprices$sale.price))$breaks
> hist(log(houseprices$sale.price), xlab = "Sale Price", axes = FALSE,
+     main = "Aranda House Price Data")
> axis(1, at = logbreaks, labels = round(exp(logbreaks), 0), tick = TRUE)
> axis(2, at = seq(0, 6), tick = TRUE)
> box()
```

Exercise 3
The `orings` data frame gives data on the damage that had occurred in US space shuttle launches prior to the disastrous Challenger launch of January 28, 1986. Only the observations in rows 1, 2, 4, 11, 13, and 18 were included in the pre-launch charts used in deciding whether to proceed with the launch.
Create a new data frame by extracting these rows from `orings`, and plot `total` incidents against `temperature` for this new data frame. Obtain a similar plot for the full data set.

Use the following to rows that hold the data that were presented in the pre-launch charts:

```
> data(orings)
> orings86 <- orings[c(1, 2, 4, 11, 13, 18), ]
```

Points are best shown with filled symbols in the first plot, and with open symbols in the second plot. (Why?)

Exercise 4
Create a data frame called `Manitoba.lakes` that contains the lake's `elevation` (in meters above sea level) and `area` (in square kilometers) as listed below. Assign the names of the lakes using the `row.names()` function. Then obtain a plot of lake area against elevation, identifying each point by the name of the lake. Because of the outlying observation, it is again best to use a logarithmic scale.
. . . . . .

Note that the data are also in the data frame `Manitoba.lakes` that is included with the *DAAG* package. Before running the code, specify

```
> data(Manitoba.lakes)
> attach(Manitoba.lakes)
```

Exercise 5
The following code extracts the lake areas from the `Manitoba.lakes` data frame and attaches the lake names to the entries of the resulting vector.

```
area.lakes <- Manitoba.lakes[[2]]
names(area.lakes) <- row.names(Manitoba.lakes)
```

Look up the help for the R function `dotchart()`. Use this function to display the data in `area.lakes`.

```
> data(Manitoba.lakes)
> area.lakes <- Manitoba.lakes[[2]]
> names(area.lakes) <- row.names(Manitoba.lakes)
> dotchart(area.lakes, pch = 16, main = "Areas of Large Manitoba Lakes",
+     xlab = "Area (in square kilometers)")
```

Exercise 9
Run the following code:

```
gender <- factor(c(rep("female", 91), rep("male", 92)))
table(gender)
gender <- factor(gender, levels=c("male", "female"))
table(gender)
gender <- factor(gender, levels=c("Male", "female")) # Note the mistake
                            # The level was "male", not "Male"
table(gender)
rm(gender)                  # Remove gender
```

Explain the output from the final `table(gender)`.

The output is

```
gender
female   male
    91     92
```

```
> table(gender)
```

```
gender
  male female
    92     91
```

```
> gender <- factor(gender, levels = c("Male", "female"))
> table(gender)
```

```
gender
  Male female
     0     91
```

```
> rm(gender)
```

*Exercise 10**
The following code uses the `for()` looping function to plot graphs that compare the relative population growth (here, by the use of a logarithmic scale) for the Australian states and territories.

```
oldpar <- par(mfrow=c(2,4))
for (i in 2:9){
plot(austpop[,1], log(austpop[, i]), xlab="Year", ylab=names(austpop)[i],
     pch=16, ylim=c(0,10))}
par(oldpar)
```

Can this be done without looping? [Hint: The answer is 'yes', although this is not an exercise for novices.]

We give the code, omitting the graphs

```
> oldpar <- par(mfrow = c(2, 4))
> sapply(2:9, function(i, df) plot(df[, 1], log(df[, i]), xlab = "Year",
+     ylab = names(df)[i], pch = 16, ylim = c(0, 10)), df = austpop)
> par(oldpar)
```

There are several subtleties here:

(i) The first argument to `sapply()` can be either a list (which is, technically, a type of vector) or a vector. Here, we have supplied the vector 2:9

(ii) The second argument is a function. Here we have supplied an inline function that has two arguments. The argument `i` takes as its values, in turn, the sucessive elements in the first argument to `sapply`

(iii) Where as here the inline function has further arguments, they area supplied as additional arguments to `sapply()`. Hence the parameter `df=austpop`.

Note that `lapply()` could be used in place of `sapply()`.