Solutions for 1.5 Exercises

1. **Effect of 'the pill' on blood pressure.**

   (a) Is there a difference in blood pressure between users and non users of the drug?

   (b) An observational study (as opposed to an experiment).

   (c) Whether they are a user or not (categorical factor) and blood pressure (continuous variable categorised into ordinal categories).

   (d) Insufficient evidence to conclude that there is a significant difference between users and non-users.

   (e) Provide the actual blood pressure, rather than ranges. Include more factors that may have an impact on blood pressure.

```
> bp <- data.frame(class = c("085-090", "090-095", "095-100", "100-105",
+     "105-110", "110-115", "115-120", "120-125", "125-130", "130-135",
+     "135-140", "140-145", "145-150", "150-155", "155-160"), nonusers = c(1,
+     1, 5, 11, 11, 17, 18, 11, 9, 7, 4, 2, 2, 1, 0), users = c(0,
+     0, 4, 5, 10, 15, 17, 14, 12, 10, 5, 4, 2, 1, 1))

> par(mfrow = c(2, 1))
> plot(nonusers ~ class, data = bp)
> plot(users ~ class, data = bp)
```

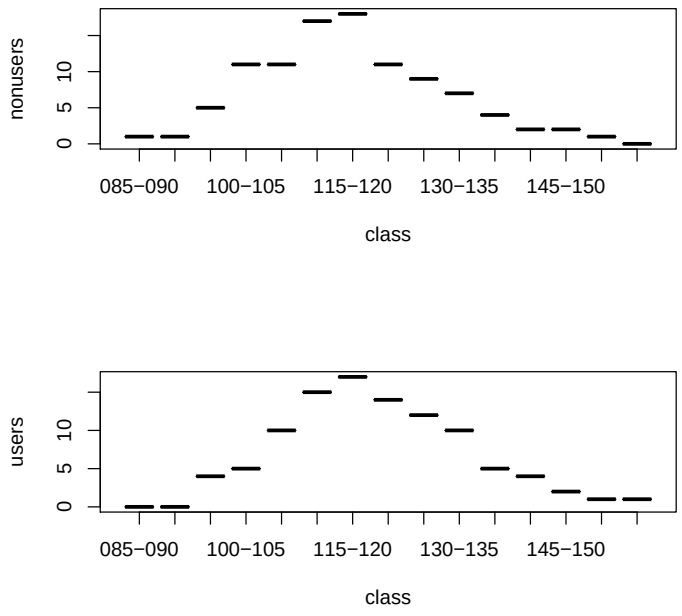


Figure 1: Non-users and users against class

2. **Pulse rate and exercise.**

```
> pulse <- read.table("../data/ms212.txt", header = TRUE)
```

(a) Example: level of other exercise.

```
> tapply(pulse$Pulse1, pulse$Exercise, mean)
```

```
       1        2        3
68.64286 75.68966 78.35135
```

```
> tapply(pulse$Pulse1, pulse$Exercise, sd)
```

```
       1        2        3
12.68923 14.09268 11.45818
```

The `tapply` function applies the function specified (in this case the mean and standard deviation) to the first argument (pulse1) for each level of the second argument (exercise).

It appears that a high level of exercise is associated with a lower pulse rate.

The variability (as measured by the standard deviation) looks similar for the three levels of exercise.

Example: height

```
> plot(pulse$Pulse1 ~ pulse$Height)
```

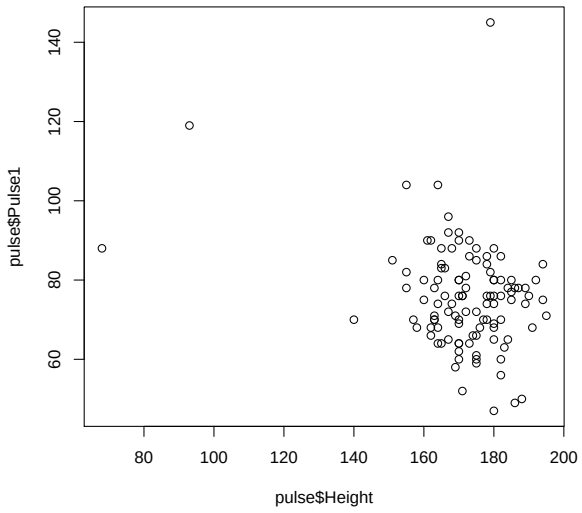


Figure 2: Scatterplot of height and pulse

There is no obvious effect of height on pulse.

(b) `> boxplot(pulse$Pulse2 ~ pulse$Ran, horizontal = TRUE)`

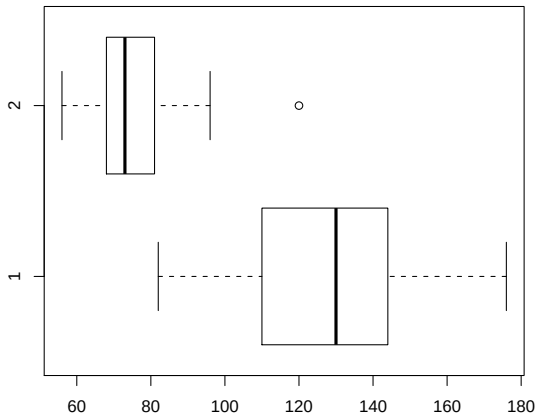


Figure 3: Boxplot of pulse

We can see that running increases the pulse.

It is also useful to examine the change in pulse.

```
> pulse$Pulsechange <- pulse$Pulse2 - pulse$Pulse1
> boxplot(pulse$Pulsechange ~ pulse$Ran, horizontal = TRUE)
```

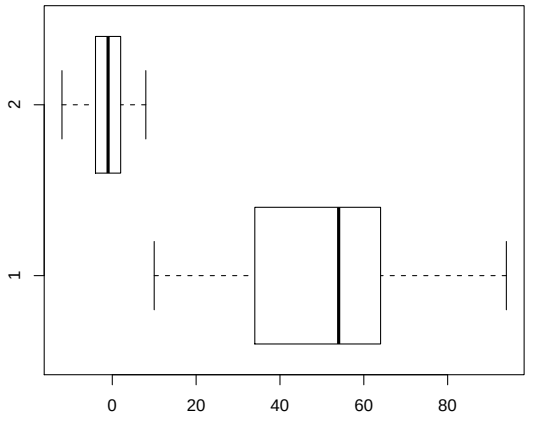


Figure 4: Boxplot of change in pulse

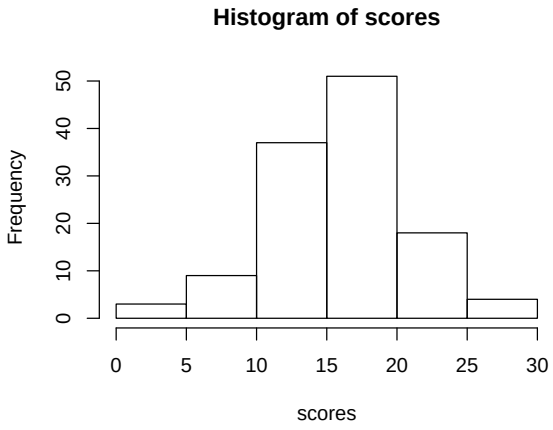
A similar picture emerges.

3. **Test results.**

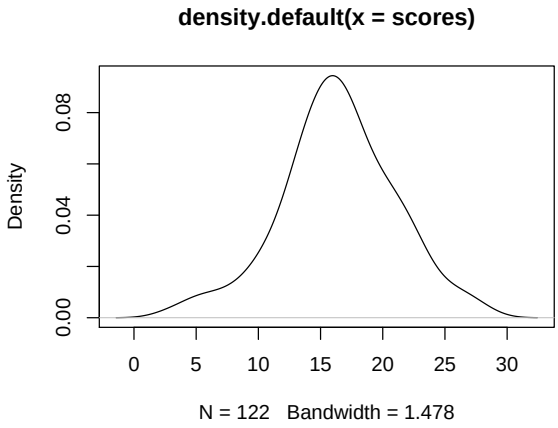
```
> scores <- c(16, 14, 15, 16, 14, 20, 18, 16, 13, 17, 16, 13, 18,
+     16, 22, 20, 24, 18, 16, 20, 26, 19, 14, 17, 14, 15, 19, 12,
+     11, 16, 19, 13, 13, 18, 8, 16, 28, 16, 27, 17, 12, 15, 7,
+     12, 22, 20, 16, 10, 8, 16, 13, 14, 14, 16, 18, 15, 21, 23,
+     16, 5, 14, 23, 17, 15, 14, 22, 20, 22, 13, 20, 18, 13, 14,
+     21, 14, 18, 10, 20, 24, 17, 21, 15, 18, 12, 23, 17, 10, 15,
+     11, 5, 16, 19, 22, 10, 15, 17, 13, 23, 20, 3, 18, 15, 22,
+     12, 9, 20, 16, 17, 17, 16, 21, 18, 11, 14, 6, 21, 25, 18,
+     26, 18, 18, 15)
> summary(scores)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.00   14.00   16.00   16.39   19.75   28.00

> hist(scores)
```




```
> plot(density(scores))
```



4. **Soil water evaporation.**

```
> soil <- data.frame(air.speed = c(0.5, 0.5, 0.5, 1, 1, 1, 1.5,
+     1.5, 1.5, 2, 2, 2, 2.5, 2.5, 2.5), evaporation = c(5.39,
+     4.43, 5.5, 7.7, 6.2, 6.14, 5.62, 6.12, 7.2, 6.88, 7.73, 6.01,
+     5.1, 7.29, 7.28))
> cor(soil$evaporation, soil$air.speed)

[1] 0.451037

> plot(evaporation ~ air.speed, data = soil)
```

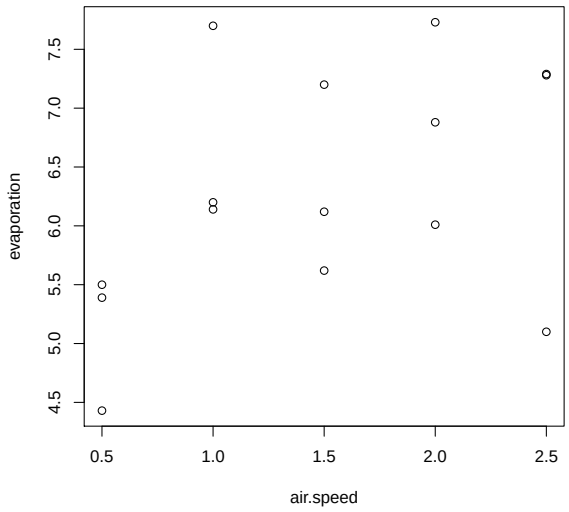


Figure 5: Scatterplot of soil data

There appears to be an increase in evaporation for increasing airspeed but there is a lot of variability and the correlation is modest.

It is possible to add features to the plot, such as a fitted line, using a command like `abline`. However in this case, while low air speed is associated with low evaporation, there is no obvious trend for higher values.

```
> plot(evaporation ~ air.speed, data = soil)
> abline(lm(soil$evaporation ~ soil$air.speed))
```

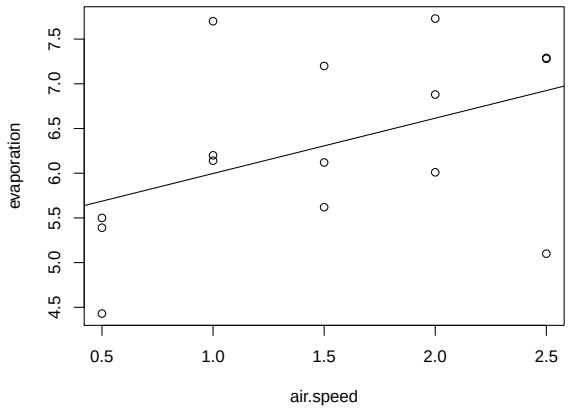


Figure 6: Scatterplot of soil data with fitted line

The function `lm` fits a linear model to the data.

5. **Melon yields.**

```
> melons <- data.frame(variety = rep(c("A", "B", "C", "D"), 6),
+     yield = c(25, 40, 18, 28, 17, 35, 23, 29, 26, 32, 26, 33,
+         16, 37, 15, 32, 22, 43, 11, 30, 16, 37, 24, 28))
> melons
```

```
   variety yield
1        A    25
2        B    40
3        C    18
4        D    28
5        A    17
6        B    35
7        C    23
8        D    29
9        A    26
10       B    32
11       C    26
12       D    33
13       A    16
14       B    37
15       C    15
16       D    32
17       A    22
18       B    43
19       C    11
20       D    30
21       A    16
22       B    37
23       C    24
24       D    28
```

```
> plot(melons)
```

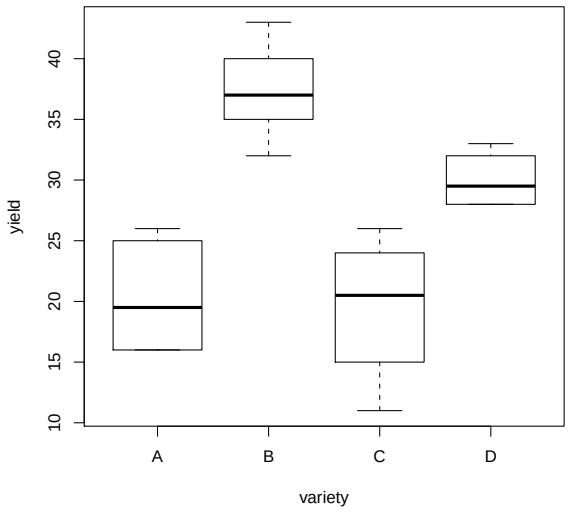


Figure 7: Plot of melon data

This plot suggests some strong differences between the melon varieties in terms of both mean yield and variation in yield.

6. **Tree inventory data.**

```
> ehc <- read.csv("../data/ehc.csv")
```

```
> plot(height.m ~ dbh.cm, data = ehc)
```

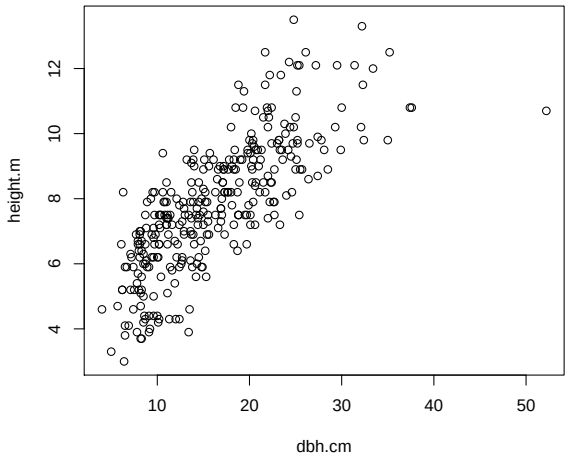


Figure 8: Plot of tree height and diameter

This plot shows a trend between dbh and height, but we are also interested in whether this differs for each species. One way of doing this is to use the `as.character` function in R, which plots the first character of each species label, as follows:

```
> plot(height.m ~ dbh.cm, pch = as.character(species), data = ehc)
```

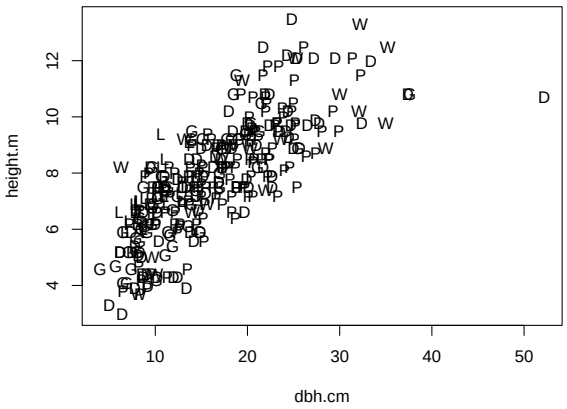


Figure 9: Plot of tree height and diameter by species

But it would be clearer if we could see a separate plot for each species, which can be achieved using the plotting function `xyplot`. This requires the lattice package, which you may need to install and load.

```
> print(xyplot(height.m ~ dbh.cm | species, data = ehc))
```

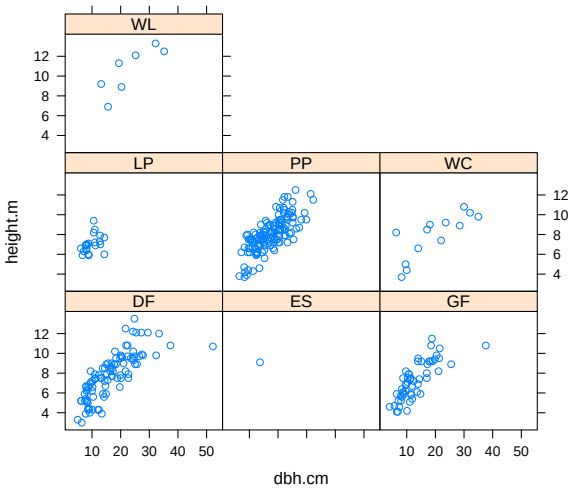


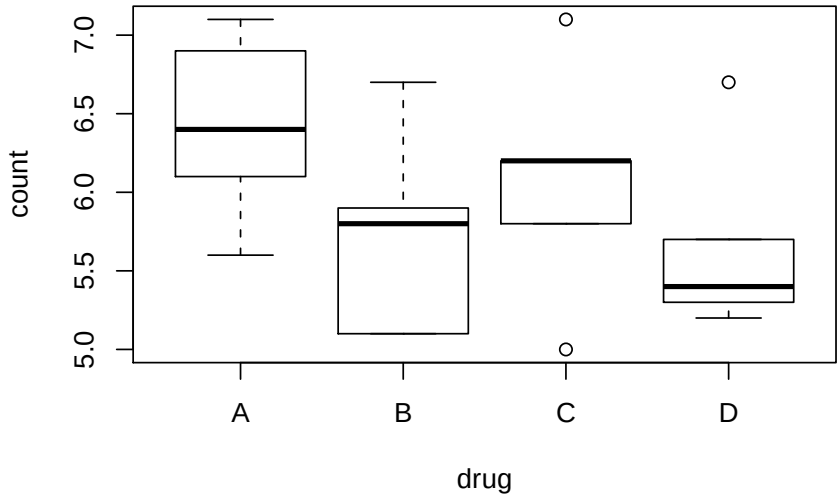
Figure 10: Plot of tree height and diameter by species

7. **Lymphocyte counts.**

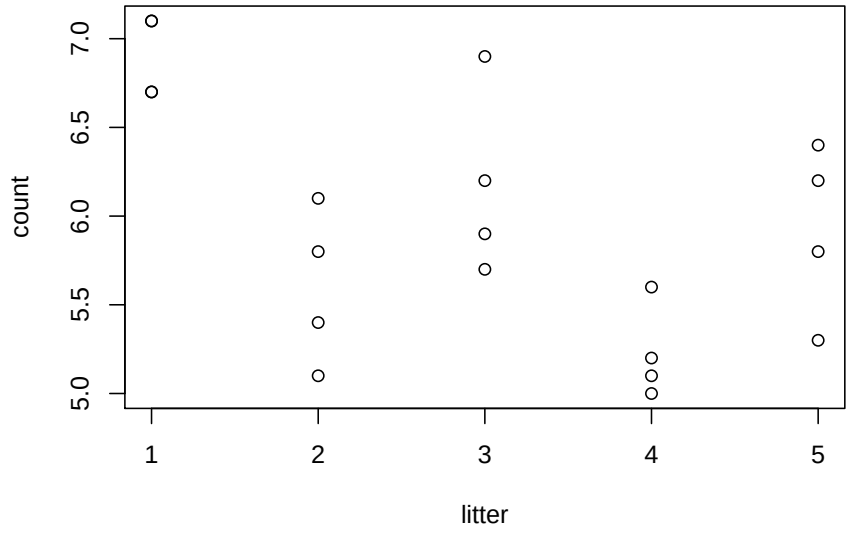
```
> count <- c(7.1, 6.7, 7.1, 6.7, 6.1, 5.1, 5.8, 5.4, 6.9, 5.9,
+     6.2, 5.7, 5.6, 5.1, 5, 5.2, 6.4, 5.8, 6.2, 5.3)
> litter <- rep(1:5, each = 4)
> drug <- rep(c("A", "B", "C", "D"), 5)
> lymph <- data.frame(count, litter, drug)
> lymph
```

	count	litter	drug
1	7.1	1	A
2	6.7	1	B
3	7.1	1	C
4	6.7	1	D
5	6.1	2	A
6	5.1	2	B
7	5.8	2	C
8	5.4	2	D
9	6.9	3	A
10	5.9	3	B
11	6.2	3	C
12	5.7	3	D
13	5.6	4	A
14	5.1	4	B
15	5.0	4	C
16	5.2	4	D
17	6.4	5	A
18	5.8	5	B
19	6.2	5	C
20	5.3	5	D

```
> plot(count ~ drug, data = lymph)
```



```
> plot(count ~ litter, data = lymph)
```



The effect of drugs on lymphocyte counts is of main interest in this experiment, and so the plot of counts vs drugs is more informative. However, a boxplot (the default chosen by R) is not the most useful plot here, as it gives a 5-number summary, and there are only 5 observations per drug. A scatterplot which shows the individual values would be better. The plot of counts vs litter has some value as it illustrates the overall differences between litters. In the analysis of the data, it would be important to take into account litter effects when comparing drugs.