Assignment 2

Solutions

Task 1:

Load in the data and convert each column:

```
# Load the data
s50 <- read.table("s50_1995.txt", header = TRUE)

# Convert each column to ordered factors with appropriate levels
s50$alcohol <- factor(s50$alcohol,
                      levels = 1:5,
                      labels = c("not", "once or twice a year", "once a month",
                                 "once a week", "more than once a week"),
                      ordered = TRUE)
s50$drugs <- factor(s50$drugs,
                      levels = 1:4,
                      labels = c("not", "once", "occasional", "regular"),
                      ordered = TRUE)
s50$smoke <- factor(s50$smoke,
                      levels = 1:3,
                      labels = c("not", "occasional", "regular"),
                      ordered = TRUE)
s50$sport <- factor(s50$sport,
                      levels = 1:2,
                      labels = c("not regular", "regular"),
                      ordered = TRUE)

# Show the structure of the dataset
str(s50)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ alcohol: Ord.factor w/ 5 levels "not"<"once or twice a year"<..: 3 2 2 2 3 4 4 4 2 4 ...
##  $ drugs  : Ord.factor w/ 4 levels "not"<"once"<"occasional"<..: 1 2 1 1 1 1 3 3 1 1 ...
##  $ smoke  : Ord.factor w/ 3 levels "not"<"occasional"<..: 2 3 1 1 1 1 1 3 1 1 ...
##  $ sport  : Ord.factor w/ 2 levels "not regular"<..: 2 1 1 2 2 2 1 2 2 2 ...
```

We can now see that the dataframe has been modified as required.

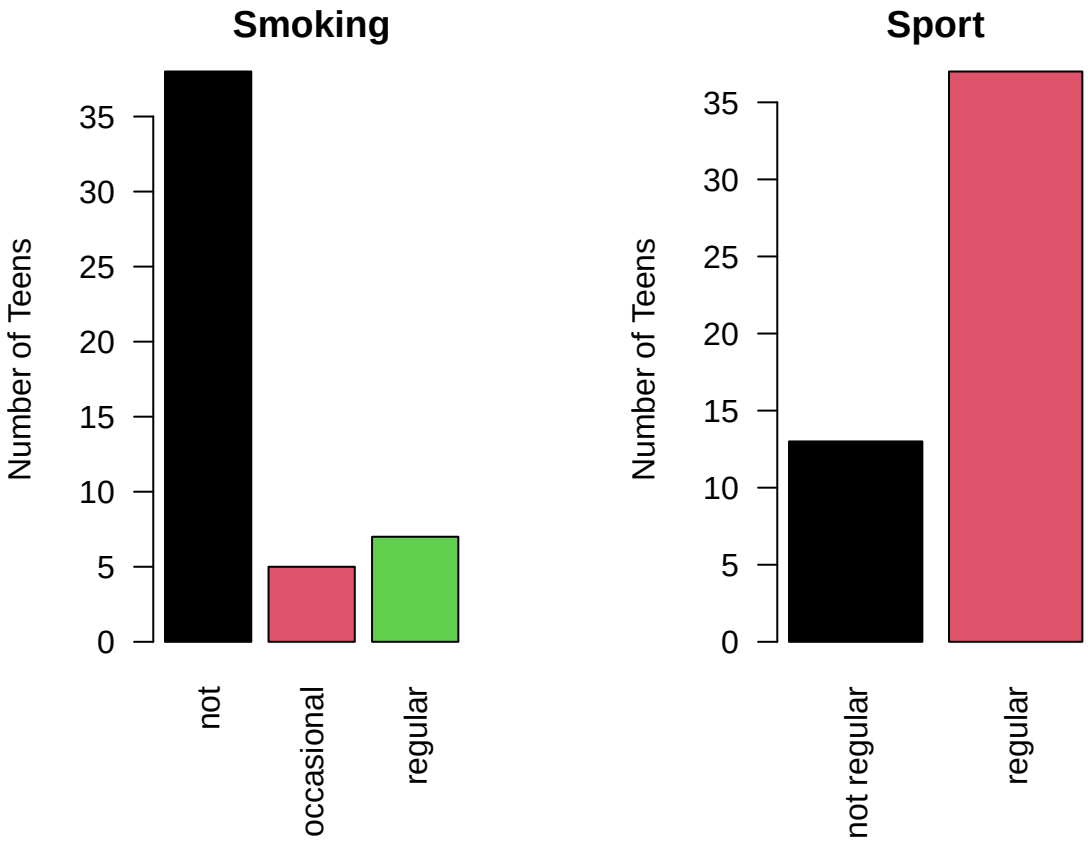
Task 2:

```
# Set figure margins:
par(mfrow=c(1, 2),
    mar=c(5, 4, 2.5, 4)) # Leave space for titles under plot

# Create first barplot:
```

```
barplot(height=table(s50$smoke),
        ylab='Number of Teens',
        main='Smoking',
        col=1:3,
        las=2)

# Create second plot:
barplot(height=table(s50$sport),
        ylab='Number of Teens',
        main='Sport',
        col=1:3,
        las=2)
```



We can see that the majority of students don't smoke and also the majority of students play sport regularly. Of those students who smoke, slightly more smoke regularly than occassionally.

We can't say anything about the *relationship* between these two variables based on these two plots.

Task 3:

```
sum(table(s50$smoke)[c("occasional", "regular")])/nrow(s50)
```

```
## [1] 0.24
```

The proportion of students who smoke at least occassionally is 0.24.

```r
# Find the number:
num <- table(s50$smoke, s50$sport)[c("occasional", "regular"),"regular"]

# Calculate the proportion
sum(num)/nrow(s50)
```

```
## [1] 0.18
```

The proportion of pupils who regularly practiced sport and smoke at least occasionally is 0.18.

Task 4:

We need to assign the class `s50survey` to the object we have created in Task 1, and then write a summary method for this class (which by definition *must* be called `summary.s50survey`). Then we employ this method and test it by running `summary(obj)` where `obj` is the name of the dataset with the assigned `s50survey` class.

```r
# Assign the class 's50survey':
class(s50) <- "s50survey"

# Write the summary method:
summary.s50survey <- function(x){
  lapply(x, function(y) table(y, dnn = NULL) / length(y))
}

# Test the method on the class instance:
summary(s50)
```

```
## $alcohol
##                   not  once or twice a year          once a month
##                  0.10                  0.32                  0.24
##           once a week more than once a week
##                  0.28                  0.06
##
## $drugs
##        not       once occasional    regular
##       0.72       0.12       0.14       0.02
##
## $smoke
##        not occasional    regular
##       0.76       0.10       0.14
##
## $sport
## not regular     regular
##        0.26        0.74
```

Task 5:

```r
summary(s50)$drugs["not"]
```

```
##  not
## 0.72
```

The proportion of pupils who did not use cannabis is 0.72.

Task 6:

Follow up data on the same students has been collected also in 1997. Reading in the file, converting each column to an ordered factor as before, and assigning the class `s50survey` to this dataset as well:

```
# Load the data
s50.1997 <- read.table("s50_1997.txt", header = TRUE)

# Convert each column to ordered factors with appropriate levels
s50.1997$alcohol <- factor(s50.1997$alcohol,
                    levels = 1:5,
                    labels = c("not", "once or twice a year", "once a month", "once a week", "more than
                    ordered = TRUE)
s50.1997$drugs <- factor(s50.1997$drugs,
                    levels = 1:4,
                    labels = c("not", "once", "occasional", "regular"),
                    ordered = TRUE)
s50.1997$smoke <- factor(s50.1997$smoke,
                    levels = 1:3,
                    labels = c("not", "occasional", "regular"),
                    ordered = TRUE)
s50.1997$sport <- factor(s50.1997$sport,
                    levels = 1:2,
                    labels = c("not regular", "regular"),
                    ordered = TRUE)

# Assign the object to the class s50survey
class(s50.1997) <- "s50survey"

# Show the structure of the dataset
str(s50.1997)
```

```
## List of 4
##  $ alcohol: Ord.factor w/ 5 levels "not"<"once or twice a year"<..: 3 2 3 2 4 4 3 4 2 4 ...
##  $ drugs  : Ord.factor w/ 4 levels "not"<"once"<"occasional"<..: 1 3 1 1 3 1 2 3 1 1 ...
##  $ smoke  : Ord.factor w/ 3 levels "not"<"occasional"<..: 1 3 1 1 1 3 3 3 1 2 ...
##  $ sport  : Ord.factor w/ 2 levels "not regular"<..: 1 1 1 1 2 2 2 2 1 2 ...
##  - attr(*, "row.names")= int [1:50] 1 2 3 4 5 6 7 8 9 10 ...
##  - attr(*, "class")= chr "s50survey"
```

Testing the `summary` S3 method on this new dataset:

```
summary(s50.1997)
```

```
## $alcohol
##                   not  once or twice a year          once a month
##                  0.02                  0.18                  0.34
##           once a week more than once a week
##                  0.34                  0.12
##
## $drugs
##        not       once occasional    regular
##       0.52       0.14       0.34       0.00
##
## $smoke
##        not occasional    regular
```

```
##       0.62       0.04       0.34
##
## $sport
## not regular     regular
##        0.62        0.38
```

Task 7:

Did the proportion of students practising sport regularly increased or decreased with respect to the 1995 data?

```
# Finding the proportion in 1997:
summary(s50.1997)$sport["regular"]
```

```
## regular
##    0.38
```

```
# And the proportion from 1995:
summary(s50)$sport["regular"]
```

```
## regular
##    0.74
```

We can see that the proportions of students practising sport regularly decreased from 0.74 to 0.38 between 1995 and 1997.