Analysis of Substance Use and Sporting Behavior among Pupils in a School in the West of Scotland

Clementine Surya

1. Load dataset

In the code below, I load the dataset of `s50_1995.txt` and convert each column to an ordered factor with appropriate labels.

```
df_1995 <- read.csv("s50_1995.txt", sep="")
df_1995$alcohol <- factor(df_1995$alcohol, levels = c("1", "2", "3","4","5"), labels = c("not", "once or twi
ce a year", "once a month","once a week","more than once a week"), ordered = TRUE)
df_1995$drugs <- factor(df_1995$drugs, levels = c("1", "2", "3","4"), labels = c("not", "tried once","occasi
onal","regular"),ordered = TRUE)
df_1995$smoke <- factor(df_1995$smoke, levels = c("1", "2", "3"),labels = c("not", "occasional","regular"),
ordered = TRUE)
df_1995$sport <- factor(df_1995$sport, levels = c("1", "2"), labels = c("not regular","regular"), ordered =
TRUE)
```

Then, I display the structure of the dataset.

```
str(df_1995)
```

```
## '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"<"tried once"<..: 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 ...
```

2. Create two suitable graphs with labels, colours. One illustrating the variable smoke and the other illustrating the variable sport. Put the two plots next to each other on the same page. Comment on the resulting plots.

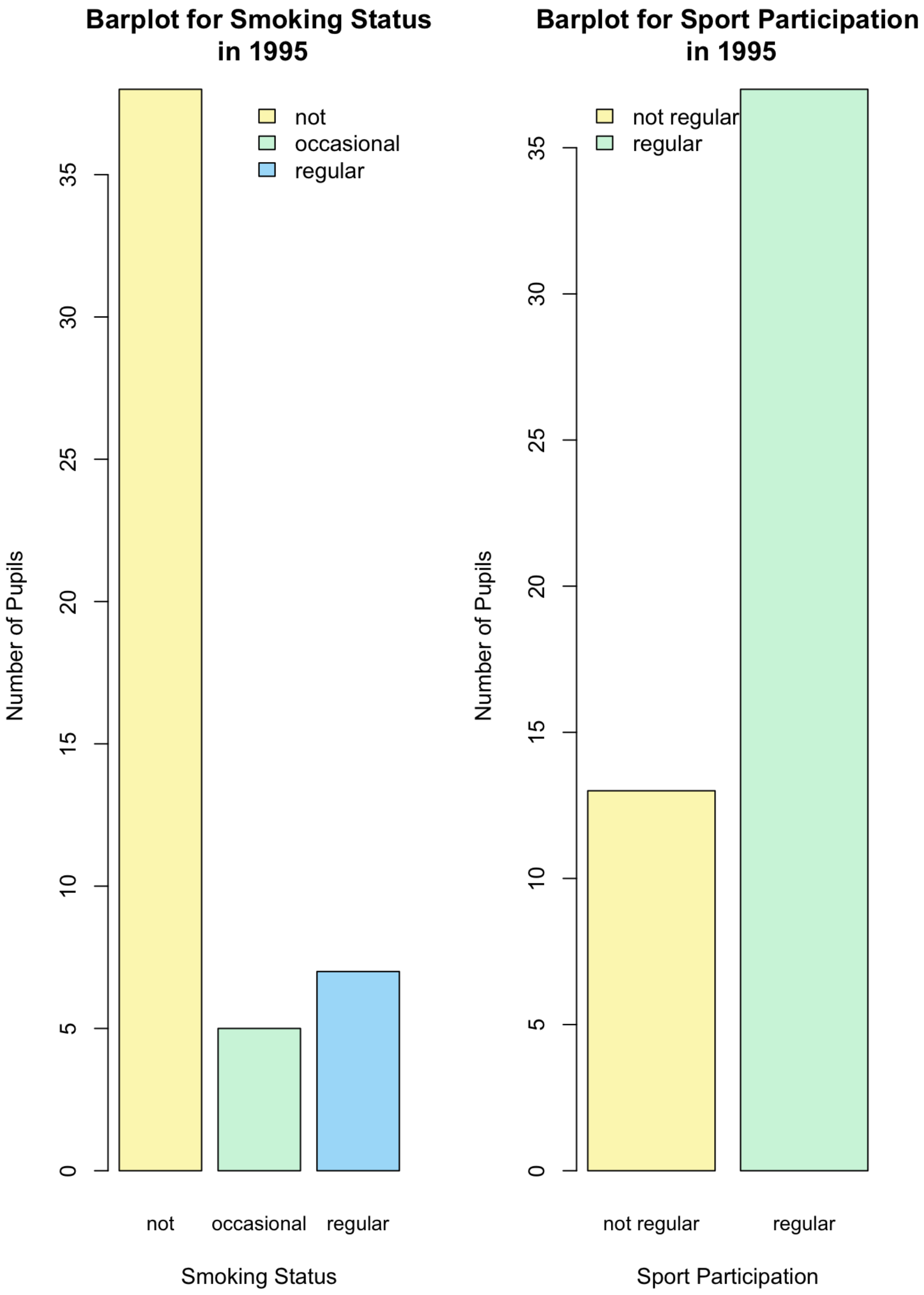
Here, I create 2 plots illustrating the variable smoke and the variable sport.

```r
#First, I create table for smoke and sport to count the number of student in each category, these will be used in creating the plots
table_smoke = table(df_1995$smoke)
table_sport = table(df_1995$sport)

#Below code is to put the two plots next to each other
par(mfrow=c(1,2))

#Create the 1st plot, Smoking Status Plot
barplot(table_smoke,
        main = 'Barplot for Smoking Status\n in 1995',
        xlab = 'Smoking Status',
        ylab = 'Number of Pupils',
        col = c("#FCF6BD","#D0F4DE","#A9DEF9"),
        cex.names = 0.9)
legend("topright",
       legend = c("not", "occasional","regular"),
       fill = c("#FCF6BD","#D0F4DE","#A9DEF9"),
       bty="n")

#Create the 2nd plot, Sport Participation Plot
barplot(table_sport,
        main = 'Barplot for Sport Participation\n in 1995',
        xlab = 'Sport Participation',
        ylab = 'Number of Pupils',
        col = c("#FCF6BD","#D0F4DE"),
        cex.names = 0.9)
legend("topleft",
       legend = c("not regular", "regular"),
       fill = c("#FCF6BD","#D0F4DE"),
       bty="n")
```



It can be seen above in the Barplot of Smoking Status that in the Year 1995, the proportion of pupils that did not smoke was the biggest followed by pupils that smoked regularly and occasionally while in the Barplot of Sport Participation, also in the same Year, the proportion of pupils who did sport regularly was bigger compared to pupils that did not do regular sport.

3. Produce some code to answer the following questions:

- What is the proportion of pupils who smoke at least occasionally?
  The proportion of pupils who smoke at least occasionally is 0.24 (24%).

```
smoke_atleast_occasionally = (sum(df_1995$smoke == "occasional") + sum(df_1995$smoke == "regular"))
proportion_smoke_atleast_ocasionally = smoke_atleast_occasionally/nrow(df_1995)
print(proportion_smoke_atleast_ocasionally)
```

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

- What is the proportion of pupils who regularly practiced sport and smoke at least occasionally?
  The proportion of pupils who regularly practiced sport and smoke at least occasionally is 0.18 (18%).

```
#For this question, I first show the proportion table of variable smoke and sport
table_smoke_sport = table(df_1995$smoke, df_1995$sport)
names(dimnames(table_smoke_sport)) <- c("Smoke", "Sport")
table_smoke_sport
```

```
##             Sport
## Smoke        not regular regular
##   not                 10      28
##   occasional           1       4
##   regular              2       5
```

```
prop_table = prop.table(table_smoke_sport)
names(dimnames(prop_table)) <- c("Smoke", "Sport")
prop_table
```

```
##             Sport
## Smoke        not regular regular
##   not               0.20    0.56
##   occasional        0.02    0.08
##   regular           0.04    0.10
```

```
#Here, I calculate the proportion of pupils who regularly practiced sport and smoke at least occasionally:
prop_table[2, 2] + prop_table[3, 2]
```

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

4. We would like to be able to summarise such data sets as new data arrive. For this reason, we want to turn the object containing the data into an S3 class called s50survey and write a summary method that will show the proportion of students for every level of each variable. Test your function on the s50_1995.txt data.

In below code, I turn the object containint the data into an S3 class called s50survey.

```
class(df_1995) <- 's50survey'
```

Here, I write summary method that will show the proportion of students for every level of each variable.

```
summary.s50survey <- function(obj) {
  cat('Proportion of students on Alcohol Consumption: \n')
  table_alcohol = table(obj$alcohol)
  prop_alcohol = prop.table(table_alcohol)
  print(prop_alcohol)
  cat('Proportion of students on Cannabis Use: \n')
  table_drugs = table(obj$drugs)
  prop_drugs = prop.table(table_drugs)
  print(prop_drugs)
  cat('Proportion of students on Smoking Status: \n')
  table_smoke = table(obj$smoke)
  prop_smoke = prop.table(table_smoke)
  print(prop_smoke)
  cat('Proportion of students on Sport Participation: \n')
  table_sport = table(obj$sport)
  prop_sport = prop.table(table_sport)
  print(prop_sport)
}
```

And then, I test my function on data `s50_1995.txt`.

```
summary(df_1995)
```

```
## Proportion of students on Alcohol Consumption:
##
##                   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
## Proportion of students on Cannabis Use:
##
##        not tried once occasional    regular
##       0.72       0.12       0.14       0.02
## Proportion of students on Smoking Status:
##
##        not occasional    regular
##       0.76       0.10       0.14
## Proportion of students on Sport Participation:
##
## not regular     regular
##        0.26        0.74
```

5. What is the proportion of pupils who did not use cannabis?

Based on the proportion in the table above and code that that I generate below, we can see that the proportion of pupils who did not use cannabis is 0.72 (72%).

```
table_drugs = table(df_1995$drugs)
prop_drugs = prop.table(table_drugs)
prop_drugs['not']
```

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

6. Follow up data on the same students has been collected also in 1997. Read in the file s50_1997.txt, convert each column to an ordered factor, and assign the class s50survey to this dataset as well. Test the summary S3 method on this new dataset.

Here, I read in the file `s50_1997.txt` and convert each column to an ordered factor.

```
df_1997 <- read.csv("s50_1997.txt", sep="")
df_1997$alcohol <- factor(df_1997$alcohol, levels = c("1", "2", "3","4","5"), labels = c("not", "once or twi
ce a year", "once a month","once a week","more than once a week"), ordered = TRUE)
df_1997$drugs <- factor(df_1997$drugs, levels = c("1", "2", "3","4"), labels = c("not", "tried once","occasi
onal","regular"),ordered = TRUE)
df_1997$smoke <- factor(df_1997$smoke, levels = c("1", "2", "3"),labels = c("not", "occasional","regular"),
ordered = TRUE)
df_1997$sport <- factor(df_1997$sport, levels = c("1", "2"), labels = c("not regular","regular"), ordered =
TRUE)
```

Then, I assign the class s50survey to this dataset as well.

```
class(df_1997) <- 's50survey'
```

After that, I test the summary S3 method on this new dataset.

```
summary(df_1997)
```

```
## Proportion of students on Alcohol Consumption:
## 
##                   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 
## Proportion of students on Cannabis Use:
## 
##        not tried once occasional    regular 
##       0.52       0.14       0.34       0.00 
## Proportion of students on Smoking Status:
## 
##        not occasional    regular 
##       0.62       0.04       0.34 
## Proportion of students on Sport Participation:
## 
## not regular     regular 
##        0.62        0.38
```

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

Based on the proportion table above and code that that I generate below, the proportion of students practicing sport regularly in 1995 and 1997 were 0.74 and 0.38, respectively. Thus, we can see that the proportion was decreased by 0.36 (36%) in Year 1997 with respect to 1995 data.

```
#Year 1995
table_sport_1995 = table(df_1995$sport)
prop_sport_1995 = prop.table(table_sport_1995)
prop_sport_1995['regular']
```

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

```
#Year 1997
table_sport_1997 = table(df_1997$sport)
prop_sport_1997 = prop.table(table_sport_1997)
prop_sport_1997['regular']
```

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