Assignment02

Junhao Ou & 21200066

2022-10-26

1. Load in the data. Convert each column to an ordered factor with appropriate labels. Display the structure of the dataset.

First of all, we will load the data.

```
df = read.table('s50_1995.txt',sep = " ",header = T )
```

Next, we will convert each column to an ordered factor.

```
df$alcohol = factor(df$alcohol, levels = c(1,2,3,4,5), labels = c('Alcohol Abstinenc
e', 'Mild Drinking', 'No-risk Drinking', 'Risk Drinking', 'High Risk Drinking'), orde
red = TRUE)
df$drugs  = factor(df$drugs, levels = c(1,2,3,4), labels = c('never', 'tried once',
'occasional', 'regular'), ordered = TRUE)
df$smoke = factor(df$smoke, levels = c(1,2,3), labels = c('never', 'occasional', 'reg
ular'), ordered = TRUE)
df$sport = factor(df$sport, levels = c(1,2), labels = c('not regular', 'regular'), or
dered = TRUE)
df
```

```
##               alcohol      drugs      smoke       sport
## 1    No-risk Drinking      never occasional     regular
## 2       Mild Drinking tried once    regular not regular
## 3       Mild Drinking      never      never not regular
## 4       Mild Drinking      never      never     regular
## 5    No-risk Drinking      never      never     regular
## 6       Risk Drinking      never      never     regular
## 7       Risk Drinking occasional      never not regular
## 8       Risk Drinking occasional    regular     regular
## 9       Mild Drinking      never      never     regular
## 10      Risk Drinking      never      never     regular
## 11 High Risk Drinking tried once    regular     regular
## 12 High Risk Drinking occasional    regular     regular
## 13   No-risk Drinking occasional      never not regular
## 14   No-risk Drinking      never      never not regular
## 15      Risk Drinking      never occasional     regular
## 16      Risk Drinking tried once occasional     regular
## 17      Mild Drinking      never      never not regular
## 18      Risk Drinking      never      never not regular
## 19   No-risk Drinking      never      never     regular
## 20      Mild Drinking      never      never     regular
## 21 Alcohol Abstinence      never      never     regular
## 22   No-risk Drinking      never      never not regular
## 23      Risk Drinking    regular    regular     regular
## 24   No-risk Drinking      never      never     regular
## 25   No-risk Drinking      never      never     regular
## 26      Risk Drinking occasional    regular     regular
## 27      Mild Drinking      never      never     regular
## 28      Mild Drinking      never      never     regular
## 29   No-risk Drinking tried once      never     regular
## 30 Alcohol Abstinence      never      never     regular
## 31      Risk Drinking      never      never     regular
## 32      Risk Drinking      never      never     regular
## 33   No-risk Drinking      never      never     regular
## 34      Mild Drinking      never      never     regular
## 35   No-risk Drinking tried once      never     regular
## 36      Risk Drinking      never      never     regular
## 37      Mild Drinking      never      never     regular
## 38   No-risk Drinking      never      never     regular
## 39      Mild Drinking      never      never     regular
## 40 Alcohol Abstinence      never      never     regular
## 41      Risk Drinking      never occasional     regular
## 42      Risk Drinking occasional    regular not regular
## 43      Mild Drinking      never      never     regular
## 44 High Risk Drinking occasional occasional not regular
## 45      Mild Drinking      never      never     regular
## 46      Mild Drinking      never      never not regular
## 47      Mild Drinking      never      never not regular
## 48      Mild Drinking      never      never     regular
## 49 Alcohol Abstinence      never      never not regular
## 50 Alcohol Abstinence tried once      never     regular
```

Then, we will check the structure of the dataframe.

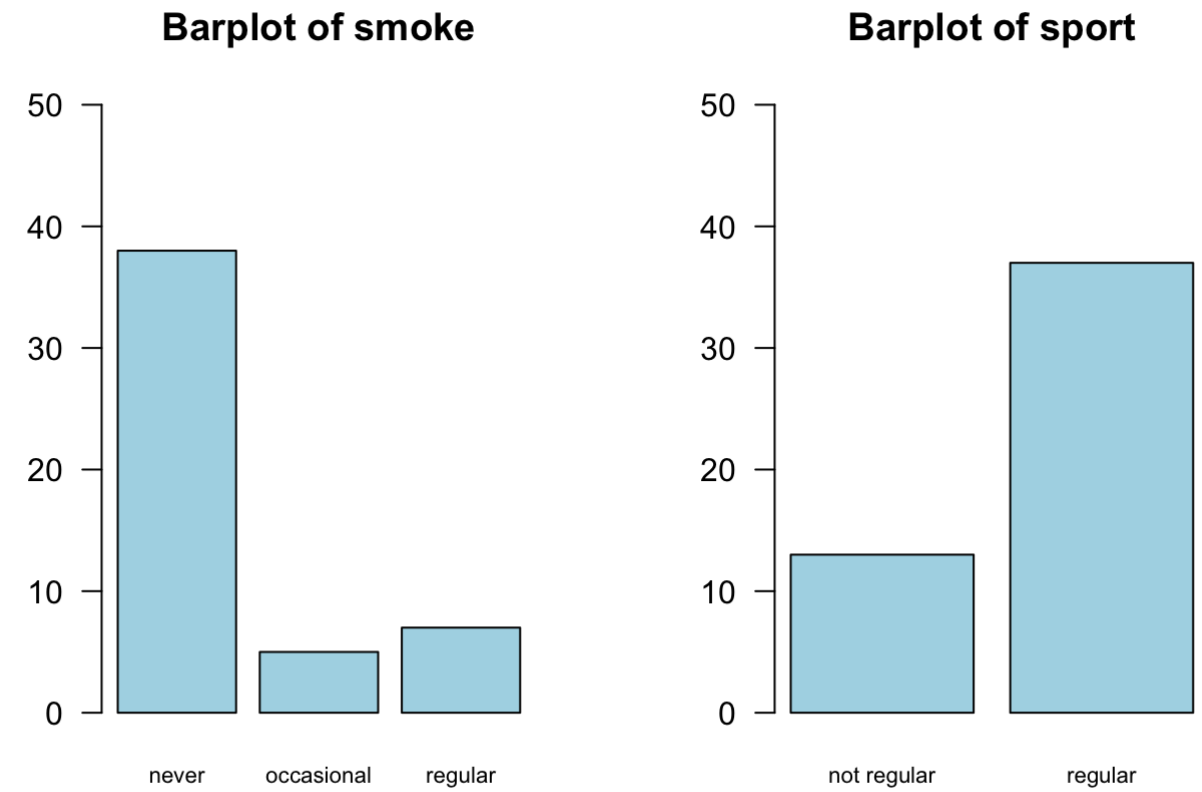
```
str(df)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ alcohol: Ord.factor w/ 5 levels "Alcohol Abstinence"<..: 3 2 2 2 3 4 4 4 2 4
...
##  $ drugs  : Ord.factor w/ 4 levels "never"<"tried once"<..: 1 2 1 1 1 1 3 3 1 1
...
##  $ smoke  : Ord.factor w/ 3 levels "never"<"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 ...
```

As we can see above, the dataframe has 50 observations of 4 variables.

2. Using base R, create two suitable graphs, with labels, colours etc., 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.

```
par(mfrow=c(1,2))
barplot(table(df$smoke), main = "Barplot of smoke", col = "lightblue", las = 1, cex.n
ames = 0.7, ylim = c(0,50))
barplot(table(df$sport), main = "Barplot of sport", col = "lightblue", las = 1, cex.n
ames = 0.7, ylim = c(0,50))
```



From the bar plots above, we can see that among the 50 investigated pupils, more than 35 had never smoked, while 5 pupils took smoking occasionally and 7 pupils took smoking regularly. When it comes to sport, around 13 pupils were reported not having regular sport, whereas 37 pupils were of regular sport.

3. Produce some code to answer the following questions:

• What is the proportion of pupils who smoke at least occasionally?

```
proportion1 = length(which(df$smoke!='never'))/length(df$smoke)
print(paste("The proportion of pupils who smoke at least occasionally is (%)", propor
tion1*100))
```

```
## [1] "The proportion of pupils who smoke at least occasionally is (%) 24"
```

• What is the proportion of pupils who regularly practiced sport and smoke at least occasionally?

```
proportion2 = length(which((df$smoke!='never') & (df$sport =='regular')))/length(df$s
moke)
print(paste("The proportion of pupils who regularly practiced sport and smoke at leas
t occasionally is (%)", proportion2*100))
```

```
## [1] "The proportion of pupils who regularly practiced sport and smoke at least occ
asionally is (%) 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.

```
#Create the class 's50survey'
class(df) <- "s50survey"

#Define a function to calculate the proportion of students for any level of label of
 each variable.
summary.s50survey = function(df) {
    for (i in 1:length(df)){
      print(table(df[i])/sum(table(df[i])))
    }
}

#Test the function on the s50_1995.txt data.
summary(df)
```

```
## alcohol
## Alcohol Abstinence      Mild Drinking   No-risk Drinking      Risk Drinking
##               0.10               0.32               0.24               0.28
## High Risk Drinking
##               0.06
## drugs
##      never tried once occasional    regular
##       0.72       0.12       0.14       0.02
## smoke
##      never occasional    regular
##       0.76       0.10       0.14
## sport
## not regular     regular
##        0.26        0.74
```

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

```
proportion3 = length(which(df$drugs == 'never'))/length(df$drugs)
print(paste("The proportion of pupils who did not use cannabi is", proportion3*100,
"%."))
```

```
## [1] "The proportion of pupils who did not use cannabi is 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.

First, we will load the data ‘s50_1997.txt’ and convert column to an ordered factor.

```
df_1997 = read.table('s50_1997.txt',sep = " ",header = T )
df_1997$alcohol = factor(df_1997$alcohol, levels = c(1,2,3,4,5), labels = c('Alcohol
 Abstinence', 'Mild Drinking', 'No-risk Drinking', 'Risk Drinking', 'High Risk Drinki
ng'), ordered = TRUE)
df_1997$drugs  = factor(df_1997$drugs, levels = c(1,2,3,4), labels = c('never', 'trie
d once', 'occasional', 'regular'), ordered = TRUE)
df_1997$smoke = factor(df_1997$smoke, levels = c(1,2,3), labels = c('never', 'occasio
nal', 'regular'), ordered = TRUE)
df_1997$sport = factor(df_1997$sport, levels = c(1,2), labels = c('not regular', 'reg
ular'), ordered = TRUE)
df_1997
```

```
##               alcohol      drugs      smoke       sport
## 1    No-risk Drinking      never      never not regular
## 2       Mild Drinking occasional    regular not regular
## 3    No-risk Drinking      never      never not regular
## 4       Mild Drinking      never      never not regular
## 5       Risk Drinking occasional      never     regular
## 6       Risk Drinking      never    regular     regular
## 7    No-risk Drinking tried once    regular     regular
## 8       Risk Drinking occasional    regular     regular
## 9       Mild Drinking      never      never not regular
## 10      Risk Drinking      never occasional     regular
## 11 High Risk Drinking tried once      never not regular
## 12 High Risk Drinking occasional    regular not regular
## 13      Mild Drinking occasional      never not regular
## 14   No-risk Drinking      never    regular not regular
## 15 High Risk Drinking occasional    regular not regular
## 16      Risk Drinking tried once    regular not regular
## 17      Risk Drinking tried once occasional not regular
## 18   No-risk Drinking tried once      never not regular
## 19 High Risk Drinking occasional    regular not regular
## 20   No-risk Drinking      never      never     regular
## 21   No-risk Drinking      never      never     regular
## 22   No-risk Drinking      never      never     regular
## 23      Mild Drinking occasional    regular not regular
## 24   No-risk Drinking tried once      never     regular
## 25      Risk Drinking      never      never not regular
## 26   No-risk Drinking occasional    regular not regular
## 27   No-risk Drinking      never      never not regular
## 28      Risk Drinking      never      never     regular
## 29   No-risk Drinking      never      never     regular
## 30      Risk Drinking occasional    regular not regular
## 31      Risk Drinking      never      never     regular
## 32      Risk Drinking      never      never not regular
## 33   No-risk Drinking      never    regular     regular
## 34      Mild Drinking      never      never not regular
## 35      Risk Drinking occasional      never not regular
## 36      Risk Drinking occasional    regular not regular
## 37   No-risk Drinking      never      never     regular
## 38      Risk Drinking occasional      never not regular
## 39   No-risk Drinking      never      never     regular
## 40 Alcohol Abstinence      never      never     regular
## 41      Risk Drinking occasional    regular not regular
## 42 High Risk Drinking occasional    regular not regular
## 43      Risk Drinking tried once      never     regular
## 44 High Risk Drinking occasional      never not regular
## 45      Mild Drinking      never      never     regular
## 46      Mild Drinking      never      never not regular
## 47      Mild Drinking      never      never not regular
## 48      Risk Drinking      never      never     regular
## 49   No-risk Drinking      never      never not regular
## 50   No-risk Drinking occasional    regular not regular
```

Then, we will assign the class s50survey to this dataset and test the summary S3 method on this new dataset.

```
class(df_1997) <- "s50survey"
summary(df_1997)
```

```
## alcohol
## Alcohol Abstinence      Mild Drinking   No-risk Drinking      Risk Drinking
##               0.02               0.18               0.34               0.34
## High Risk Drinking
##               0.12
## drugs
##      never tried once occasional    regular
##       0.52       0.14       0.34       0.00
## smoke
##      never occasional    regular
##       0.62       0.04       0.34
## sport
## 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?

```
proportion4 = length(which(df$sport == 'regular'))/length(df$sport)
print(paste("The proportion of pupils practising sport regularly in 1995 is (%)", pro
portion4*100))
```

```
## [1] "The proportion of pupils practising sport regularly in 1995 is (%) 74"
```

```
proportion5 = length(which(df_1997$sport == 'regular'))/length(df_1997$sport)
print(paste("The proportion of pupils practising sport regularly in 1997 is (%)", pro
portion5*100))
```

```
## [1] "The proportion of pupils practising sport regularly in 1997 is (%) 38"
```

From the result above, we can conclude that the proportion of pupils practising sport regularly experienced a decrease from 74% in 1995 to 38% in 1997.