Substance use and sporting behaviour of pupils Analysis

Author : Vedanti Borate

Data Loading and Cleaning

Loading Data

```
#set path
  setwd("~/Documents/GitCodes/RProgramming")
#read txt file as table
   s50 <- read.table("3_s50_1995.txt", header = TRUE)
```

Factoring Dataset

```
# factoring and labeling to make s50 more readable

s50$alcohol <-
  factor(
    s50$alcohol,
    ordered = TRUE,
    levels = 1:5,
    labels = c(
      "not",
      "once or twice a year",
      "once a month",
      "once a week",
      "more than once a week"
    )
  )
s50$drugs <-
  factor(
    s50$drugs,
    ordered = TRUE,
    levels = 1:4,
    labels = c("not", "tried once", "occasional", "regular")
  )
s50$smoke <-
```

```
  factor(
    s50$smoke,
    ordered = TRUE,
    levels = 1:3,
    labels = c("not", "occasional", "regular")
  )
s50$sport <-
  factor(
    s50$sport,
    ordered = TRUE,
    levels = 1:2,
    labels = c("not regular", "regular")
  )
```

Review Dataset structure

```
#checking the structure of 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
##  $ drugs  : Ord.factor w/ 4 levels "not"<"tried once"<..: 1 2 1 1 1 1 3 3 1
##  $ smoke  : Ord.factor w/ 3 levels "not"<"occasional"<..: 2 3 1 1 1 1 1 3 1
##  $ sport  : Ord.factor w/ 2 levels "not regular"<..: 2 1 1 2 2 2 1 2 2 2 ...
```

Modified Dataset

```
#review data with labelled information
head(s50,10)
```

```
##                 alcohol      drugs      smoke       sport
## 1          once a month        not occasional     regular
## 2  once or twice a year tried once    regular not regular
## 3  once or twice a year        not        not not regular
## 4  once or twice a year        not        not     regular
## 5          once a month        not        not     regular
## 6           once a week        not        not     regular
## 7           once a week occasional        not not regular
## 8           once a week occasional    regular     regular
## 9  once or twice a year        not        not     regular
## 10          once a week        not        not     regular
```

Data Visualization

Observations:

Plots are taken in 1995 for collecting pupils data.By looking at both the plots we can see that there are max 50 students. They are divided into different categories of smoking and playing sports regularity. Both the graphs are not exactly related as there are other factors as well in the dataset.

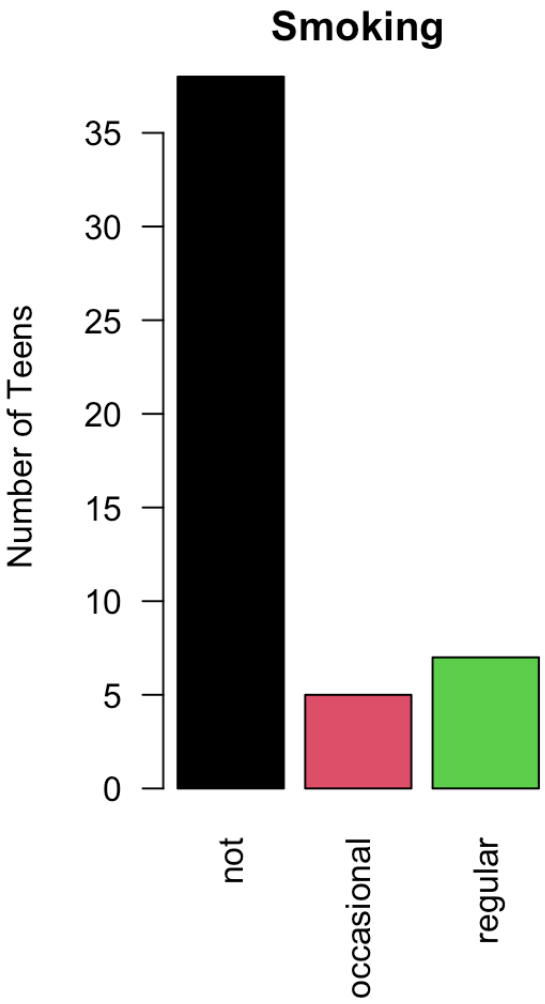
First graph is of Students smoking status. We can see that maximum students do not smoke and there are few who smoke occassional.

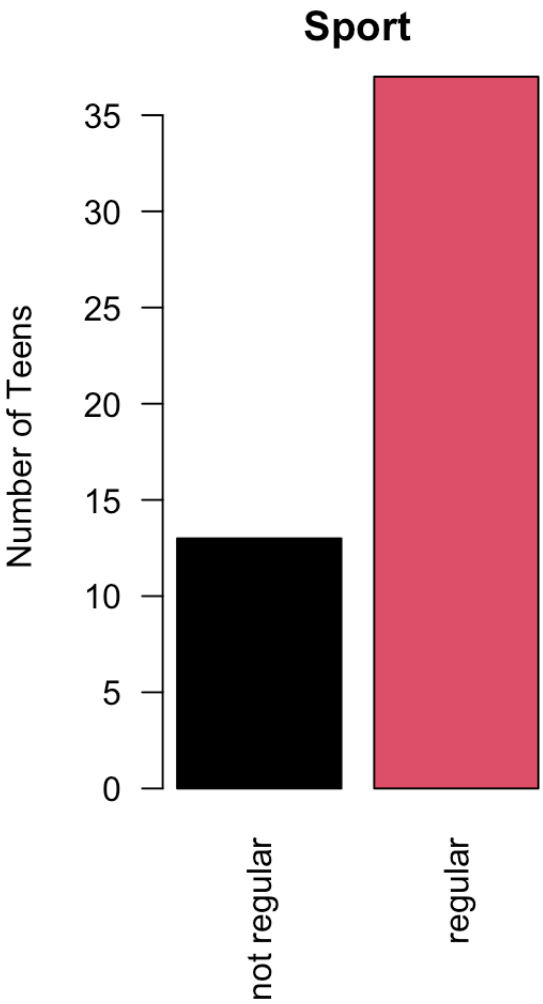
Alongside the graph, there is another graph relating to students who pratice sports. Students who pratice regular are more than non regular once.

We can't establish any relationship between these two variables based on these two plots.

```
#Set figure margins:
  par(mfrow=c(1, 2),   #No of rows and cols in graphic area
  mar=c(5, 4, 2.5, 4)) # 4 sides margins
# Create first barplot:
  barplot(height=table(s50$smoke),    #create freq table and plot that
          ylab='Number of Teens',
          main='Smoking',
          col=1:3,                   #default colors for bars
          las=2)                     # labels, parallel to x axis

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





Statistical Analysis

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

The proportion of pupils who smoke at least occasionally is **0.24**.

```
smokeandpratice = sum(s50$sport=='regular' & s50$smoke!="not") / nrow(s50)
```

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

```
summary.s50survey <- function(obj)
  {
  names <- names(obj)                                  #names of variables in survey
  list <- vector("list", length = length(names)) #empty list of same length as
  names(list) <- names                                 #list now has same names as ob
  for(i in 1:length(list))
    {                                          #loop creating proportions and adding to li
    list[[i]] <- table(obj[[i]])/length(obj[[i]])
     }
  return(list)                                         #output of summary function
```

```
  }

summary(s50)
```

```
##                   alcohol          drugs           smoke            sport
##  not                  : 5   not       :36   not       :38   not regular:13
##  once or twice a year :16   tried once: 6   occasional: 5   regular    :37
##  once a month         :12   occasional: 7   regular   : 7
##  once a week          :14   regular   : 1
##  more than once a week: 3
```

Observations: Here we find a summary of data which explains the count of each variable against the total count of pupils(or number of records in dataset).We create a class, mention a summary function for it and then we can call the print method for it.

Below are the Proportion wise details of data.

```
# 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 tried 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
```

Depending on the above proportions summary we find specifics.

```
Pdrugs = summary(s50)$drugs["not"]
```

Proportion of pupils who do not take cannabis **0.72**.

Loading Data for 1997 and following same analysis

```
#load new dataset for 1997 year
  s70 <- read.table("3_s50_1997.txt", header = TRUE)
```

Factoring Dataset

```
# factoring and labeling to make s50 more readable

s70$alcohol <-
  factor(
    s70$alcohol,
    ordered = TRUE,
    levels = 1:5,
    labels = c(
      "not",
      "once or twice a year",
      "once a month",
      "once a week",
      "more than once a week"
    )
  )
s50$drugs <-
  factor(
    s50$drugs,
    ordered = TRUE,
    levels = 1:4,
    labels = c("not", "tried once", "occasional", "regular")
  )
s70$smoke <-
  factor(
    s70$smoke,
    ordered = TRUE,
    levels = 1:3,
    labels = c("not", "occasional", "regular")
```

```
  )
s70$sport <-
  factor(
    s70$sport,
    ordered = TRUE,
    levels = 1:2,
    labels = c("not regular", "regular")
  )
```

Review Dataset structure

```
#check structure
str(s70)
```

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

1997_pupildata_proportions

```
# Assign the class 's50survey':
  class(s70) <- "s50survey"
# Write the summary method:
  summary.s70survey <- function(x){
    lapply(x, function(y) table(y, dnn = NULL) / length(y))
    }
# Test the method on the class instance:
  summary(s70)
```

```
## $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
##    1    2    3
## 0.52 0.14 0.34
##
## $smoke
##        not occasional    regular
##       0.62       0.04       0.34
```

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

Comparing Data

```
# Finding the proportion in 1997:
summary(s70)$sport["regular"]
```

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

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

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

__Observations:__From the data in both the years, we can see that students **practicing sports regularly decreased** in 1997 then in 1995.

Proportion of sports students__ (1995): 0.74__

Proportion of sports students__ (1997): 0.38__