# STAT40730 - Assignment 2

Daniel Williams 21203054

23/11/2021

# Question 1.

1. Load in the data. Convert each column to an ordered factor with appropriate labels [Hint: look at the arguments of the function factor, in particular levels and labels]. Display the structure of the dataset.

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
x <- read_table2("s50_1995.txt")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `"alcohol"` = col_double(),
##   `"drugs"` = col_double(),
##   `"smoke"` = col_double(),
##   `"sport"` = col_double()
## )
```

```
levels_alc <- c(1,2,3,4,5)
labels_alc <- c('not', 'once or twice a year', 'once a month', 'once a week', 'more than once
a week')



levels_drugs <- c(1:4)
labels_drugs <- c('not', 'tried once', 'occasional', 'regular')

levels_smoke <- c(1:3)
labels_smoke <- c('not', 'occasional', 'regular')

levels_sport <- c(1,2)
labels_sport <- c('not regular', 'regular')



x$`"alcohol"` <- factor(x$`"alcohol"`, levels = levels_alc, labels = labels_alc, ordered = TR
UE)
x$`"drugs"` <- factor(x$`"drugs"`, levels = levels_drugs, labels = labels_drugs, ordered = TR
UE)
x$`"smoke"` <- factor(x$`"smoke"`, levels = levels_smoke, labels = labels_smoke, ordered = TR
UE)
x$`"sport"` <- factor(x$`"sport"`, levels = levels_sport, labels = labels_sport, ordered = TR
UE)

str(x)
```

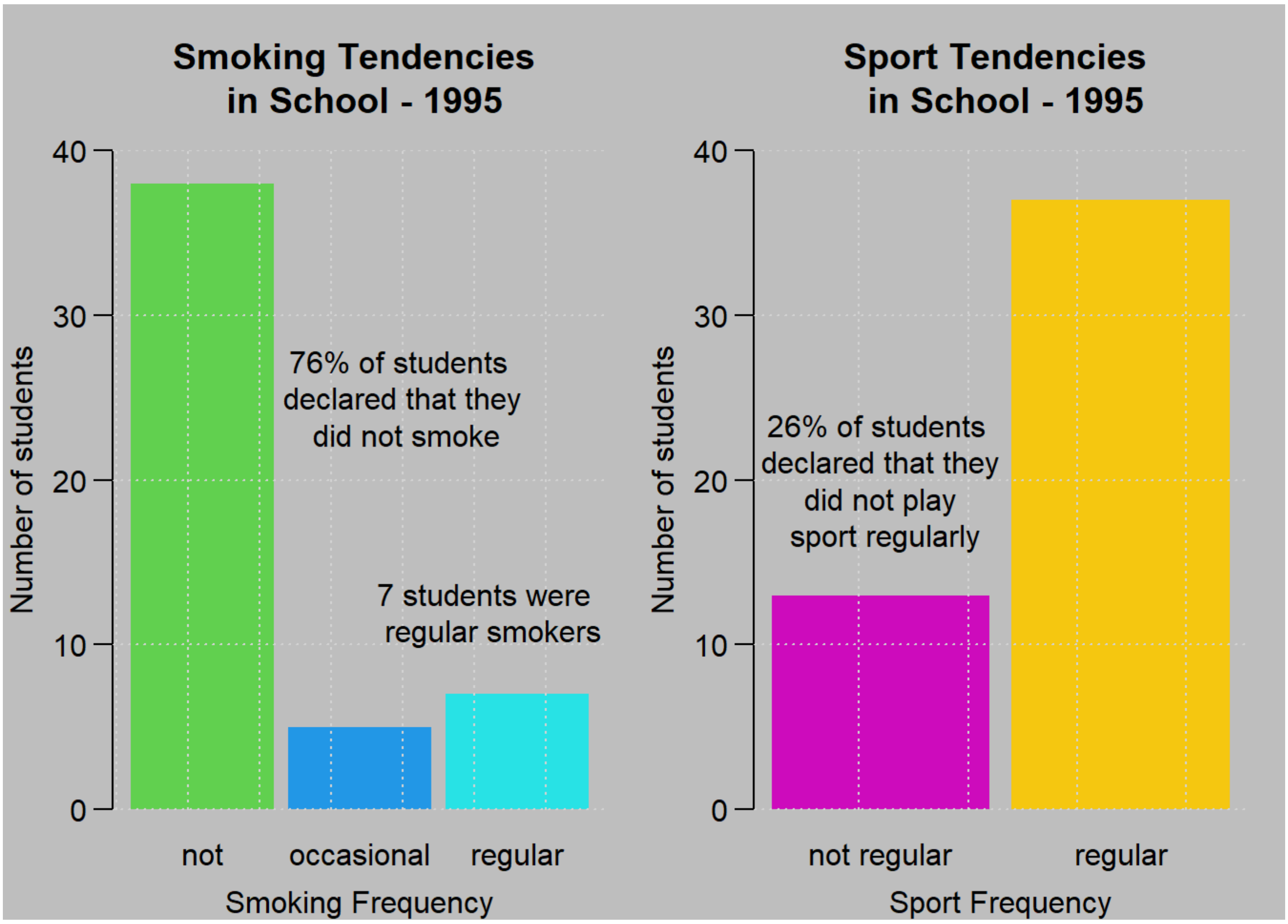
```
## spec_tbl_df [50 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ "alcohol": Ord.factor w/ 5 levels "not"<"once or twice a year"<..: 3 2 2 2 3 4 4 4 2 4
...
##  $ "drugs"  : Ord.factor w/ 4 levels "not"<"tried once"<..: 1 2 1 1 1 1 3 3 1 1 ...
##  $ "smoke"  : Ord.factor w/ 3 levels "not"<"occasional"<..: 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 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `"alcohol"` = col_double(),
##   ..   `"drugs"` = col_double(),
##   ..   `"smoke"` = col_double(),
##   ..   `"sport"` = col_double()
##   .. )
```

# Question 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(mar = c(3, 3, 4, 1), mgp = c(2,0.7, 0), las = 1, bg = 'grey', mfrow = c(1,2))
barplot(height = table(x$`"smoke"`),
        main = "Smoking Tendencies \n in School - 1995",
        xlab = "Smoking Frequency", ylab = 'Number of students',
        col = 3:5, border = NA, space = c(0.1,0.1,0.1), ylim = c(0,40))
grid()
text(2,25,"76% of students \n declared that they \n did not smoke")
text(2.6,12,"7 students were \n regular smokers")

barplot(height = table(x$`"sport"`),
        main = "Sport Tendencies \n in School - 1995",
        xlab = "Sport Frequency", ylab = 'Number of students',
        col = 6:7, border = NA, space = c(0.1,0.1), ylim = c(0,40))
grid()
text(0.6,20,"26% of students \n declared that they \n did not play \n sport regularly")
```



Comments on the plots; 76% of students declared that they did not smoke 7 students were regular smokers 26% of students declared that they did not play sport regularly

# Question 3

Produce some code to answer the following questions: • What is the proportion of pupils who smoke at least occasionally? • What is the proportion of pupils who regularly practiced sport and smoke at least occasionally?

```
#code for the first question
y <- table(x$`"smoke"`)
addmargins(y, FUN = sum)
```

```
##
##        not occasional    regular        sum
##         38          5          7         50
```

```
prop <- y / sum(y)
prop[1]
```

```
##  not
## 0.76
```

```
#code for the second question
B <- x[,-c(1,2)]
Bt <- table(B)
Prop2 <- Bt / sum(Bt)
Prop2
```

```
##             "sport"
## "smoke"      not regular regular
##   not               0.20    0.56
##   occasional        0.02    0.08
##   regular           0.04    0.10
```

76% say they do not smoke, therefore 24% say they smoke at least occasionally. Reading proportions, regular & occasionally (8%), regular & regular (10%), total 18%.

# Question 4

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.

```
j <- list(x)
class(j) <- 's50survey'

summary.s50survey <- function(G) {


  F <- as.data.frame(G[1])
  colnames(F) <- c('alcohol', 'drugs', 'smoke', 'sport')

  tot <- sum(table(F$alcohol))

  alc <- table(F$alcohol)
  print('Alcohol Proportions')
  print( alc / tot )

  drugs <- table(F$drugs)
  print('Drugs Proportions')
  print( drugs / tot )

  smoke <- table(F$smoke)
  print('Smoke Proportions')
  print( smoke / tot )

  sport <- table(F$sport)
  print('Sport Proportions')
  print( sport / tot )
}

summary.s50survey(j)
```

```
## [1] "Alcohol Proportions"
## 
##                    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 
## [1] "Drugs Proportions"
## 
##         not tried once occasional    regular 
##        0.72       0.12       0.14       0.02 
## [1] "Smoke Proportions"
## 
##         not occasional    regular 
##        0.76       0.10       0.14 
## [1] "Sport Proportions"
## 
## not regular     regular 
##        0.26        0.74 
```

# Question 5

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

```
summary.s50survey(j)
```

```
## [1] "Alcohol Proportions"
## 
##                    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 
## [1] "Drugs Proportions"
## 
##        not tried once occasional    regular 
##       0.72       0.12       0.14       0.02 
## [1] "Smoke Proportions"
## 
##        not occasional    regular 
##       0.76       0.10       0.14 
## [1] "Sport Proportions"
## 
## not regular     regular 
##        0.26        0.74
```

Did not use drugs (cannabis) - 72% from above code

# Question 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.

```
data97 <- read_table2("s50_1997.txt")
```

```
## 
## -- Column specification --------------------------------------------------------
## cols(
##   `"alcohol"` = col_double(),
##   `"drugs"` = col_double(),
##   `"smoke"` = col_double(),
##   `"sport"` = col_double()
## )
```

```
data97$`"alcohol"` <- factor(data97$`"alcohol"`, levels = levels_alc, labels = labels_alc, ordered = TRUE)
data97$`"drugs"` <- factor(data97$`"drugs"`, levels = levels_drugs, labels = labels_drugs, ordered = TRUE)
data97$`"smoke"` <- factor(data97$`"smoke"`, levels = levels_smoke, labels = labels_smoke, ordered = TRUE)
data97$`"sport"` <- factor(data97$`"sport"`, levels = levels_sport, labels = labels_sport, ordered = TRUE)

list97 <- list(data97)
class(list97) <- 's50survey'

summary.s50survey(list97)
```

```
## [1] "Alcohol Proportions"
## 
##                   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 
## [1] "Drugs Proportions"
## 
##        not tried once occasional    regular 
##       0.52       0.14       0.34       0.00 
## [1] "Smoke Proportions"
## 
##        not occasional    regular 
##       0.62       0.04       0.34 
## [1] "Sport Proportions"
## 
## not regular     regular 
##        0.62        0.38
```

# Question 7

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

```
summary.s50survey(j)
```

```
## [1] "Alcohol Proportions"
## 
##                   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 
## [1] "Drugs Proportions"
## 
##        not tried once occasional    regular 
##       0.72       0.12       0.14       0.02 
## [1] "Smoke Proportions"
## 
##        not occasional    regular 
##       0.76       0.10       0.14 
## [1] "Sport Proportions"
## 
## not regular     regular 
##        0.26        0.74
```

```
summary.s50survey(list97)
```

```
## [1] "Alcohol Proportions"
## 
##                   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 
## [1] "Drugs Proportions"
## 
##        not tried once occasional    regular 
##       0.52       0.14       0.34       0.00 
## [1] "Smoke Proportions"
## 
##        not occasional    regular 
##       0.62       0.04       0.34 
## [1] "Sport Proportions"
## 
## not regular     regular 
##        0.62        0.38
```

1995 Data - 74% practicing sport regularly, 1997 Data - 38% practicing sport regularly.
The proportion has decreased