Assignment 2 - Data Programming with R

K.Saketh Sai Nigam 22201204

2022-11-03

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.

Loading the packages required to the question…..!

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

Loading the 1995 dataset and printing the 1st six rows of the data……..!

```
##   alcohol drugs smoke sport
## 1       3     1     2     2
## 2       2     2     3     1
## 3       2     1     1     1
## 4       2     1     1     2
## 5       3     1     1     2
## 6       4     1     1     2
```

––(Converting each column to an ordered factor with appropriate labels)––

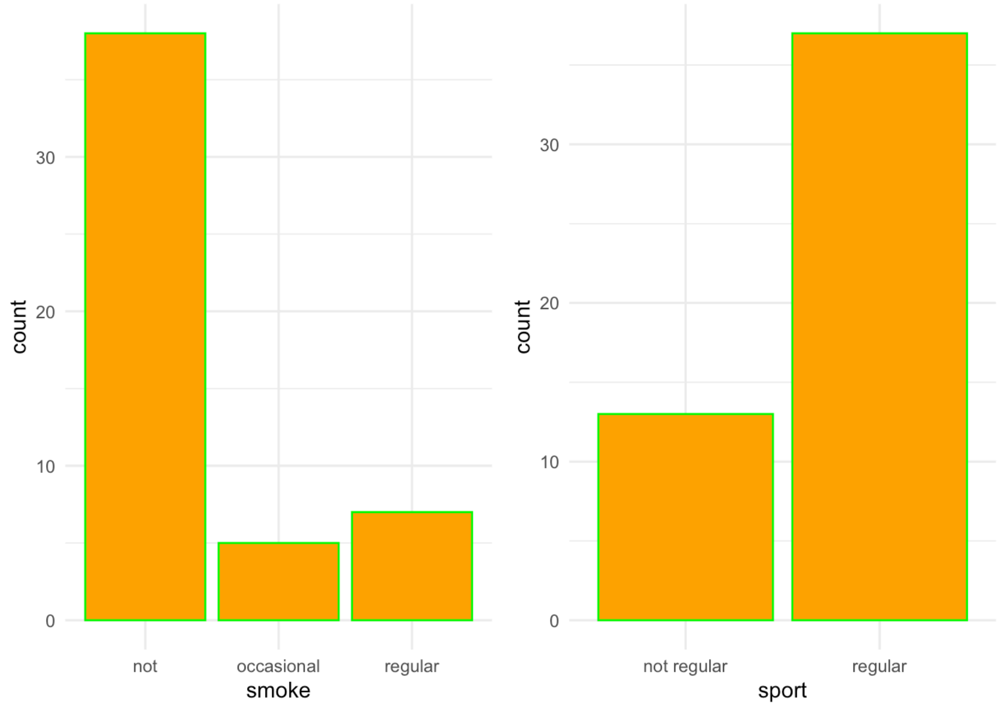
The Structure of the dataset………!

```
## '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. 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?

Plotting the 'SMOKE' n 'SPORT' data in the dataset…..!



DESCRIPTION OF PLOTS OF SMOKING STATUS AND SPORTS PARTICIPATION:

I'll now go into detail about the dataset's bar plot of the smoke column. It typically is classified into one of three categories: not smoking, occasional smoking, or regular smoking.Out of 50 children, 38 will not smoke, 5 will only smoke occasionally, and 7 will smoke frequently. This indicates that 76% of students will not smoke, 10% will smoke occasionally, and 14% will smoke frequently. Therefore, based on a comparison of the three categories, we may conclude that the majority of students will not smoke.

I'll now go into more depth about the sports column's bar plot from the dataset. It is usually divided into two categories: not regularly practiced sports and regularly practiced sports.Thirty-seven of the fifty kids will routinely practice sports, compared to thirteen who won't. This shows that 74% of students will participate in sports, whereas the remaining 26% won't do so frequently. We may therefore infer from a comparison of the two categories that most kids will practice sports activities.

---

3. Produce some code to answer the following questions:

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

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

Out of 50 children,Approximately 24% of students have smoked at least occasionally

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

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

Out of 50 children,Approximately 18% of students have practiced sports and smoked at least occasionally

---

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.

Summary of the 1995 Dataset

```
## THE STUDENT SUMMARY THAT SHOWS THE PERCENTAGE OF STUDENTS FOR  EACH LEVEL OF EACH VARIABLE:
## ********************************************************************************************
## ------------------------------------------------------------------------------------------
## 1. ALCOHOL: ALCOHOL CONSUMPTION
## ------------------------------------------------------------------------------------------
## NOT:-  0.1
## ONCE OR TWICE A YEAR:-  0.32
## ONCE A MONTH:-  0.24
## ONCE A WEEK:-  0.28
## MORE THAN ONCE A WEEK:-  0.06
## ------------------------------------------------------------------------------------------
## 2. DRUGS: CANNABIS
## ------------------------------------------------------------------------------------------
## NOT:-  0.72
## TRIED ONCE:-  0.12
## OCCASIONAL:-  0.14
## REGULAR:-  0.02
## ------------------------------------------------------------------------------------------
## 3. SMOKE: SMOKING STATUS
## ------------------------------------------------------------------------------------------
## NOT:-  0.76
## OCCASIONAL:-  0.1
## REGULAR:-  0.14
## ------------------------------------------------------------------------------------------
## 4. SPORT: SPORT PARTICIPATION
## ------------------------------------------------------------------------------------------
## NOT REGULAR:-  0.26
## REGULAR:-  0.74
```

——

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

```
## [1] 0.72
```

Out of 50 children,Approximately 72% of students have not used cannabis drugs

——

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.

Loading the 1997 dataset and printing the 1st six rows of the data……..!

```
##   alcohol drugs smoke sport
## 1       3     1     1     1
## 2       2     3     3     1
## 3       3     1     1     1
## 4       2     1     1     1
## 5       4     3     1     2
## 6       4     1     3     2
```

—–(Converting each column to an ordered factor with appropriate labels)—–

Summary of the 1997 Dataset

```
## THE STUDENT SUMMARY THAT SHOWS THE PERCENTAGE OF STUDENTS FOR  EACH LEVEL OF EACH VARIABLE:
## ********************************************************************************************
## ------------------------------------------------------------------------------------------
## 1. ALCOHOL: ALCOHOL CONSUMPTION
## ------------------------------------------------------------------------------------------
## NOT:-  0.02
## ONCE OR TWICE A YEAR:-  0.18
## ONCE A MONTH:-  0.34
## ONCE A WEEK:-  0.34
## MORE THAN ONCE A WEEK:-  0.12
## ------------------------------------------------------------------------------------------
## 2. DRUGS: CANNABIS
## ------------------------------------------------------------------------------------------
## NOT:-  0.52
## TRIED ONCE:-  0.14
## OCCASIONAL:-  0.34
## REGULAR:-  0
## ------------------------------------------------------------------------------------------
## 3. SMOKE: SMOKING STATUS
## ------------------------------------------------------------------------------------------
## NOT:-  0.62
## OCCASIONAL:-  0.04
## REGULAR:-  0.34
## ------------------------------------------------------------------------------------------
## 4. SPORT: SPORT PARTICIPATION
## ------------------------------------------------------------------------------------------
## NOT REGULAR:-  0.62
## REGULAR:-  0.38
```

---

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

```
## [1] 0.74
```

```
## [1] 0.38
```

Out of 50 adolescents, 38% of pupils participated in sport in 1997, compared to 74% of children who did so in 1995. Thus, we see that from 1995 to 1997, the percentage decreased by up to 36%.