How to approach data

SSSA - Applied Statistics - Chiara Seghieri and Costanza Tortù

2023-11-07

Preliminaries

Recall packages

Before loading packages, please install it! To install a package called "package" you can use the function install.packages("package")

Import Data

```
crime <- read.dta("~/Documents/Sant'Anna/Corso allievi/Data/crimedata.dta")
```

Have a first look at data

Inspect variables

```
dim(crime) # units x variables
```

```
## [1] 8843   32
```

```
colnames(crime) # have a look at the names of the variables
```

```
##  [1] "rowlabel"  "split"     "sex"       "yrsarea"   "resyrago"  "work2"
##  [7] "tenure1"   "livharm1"  "agegrp7"   "ethgrp2a"  "educat3"   "rural2"
## [13] "edeprivex" "wdeprivex" "IndivWgtx" "cause2m"   "walkdark"  "walkday"
## [19] "homealon"  "wburgl"    "wmugged"   "wcarstol"  "wfromcar"  "wraped"
## [25] "wattack"   "wraceatt"  "worryx"    "bcsvictim" "rubbcomm"  "vandcomm"
## [31] "poorhou"   "antisocx"
```

```
table(crime$bcsvictim, crime$sex)
```

Look at the joint distribution of sex and having experienced at least one crime in the previous year

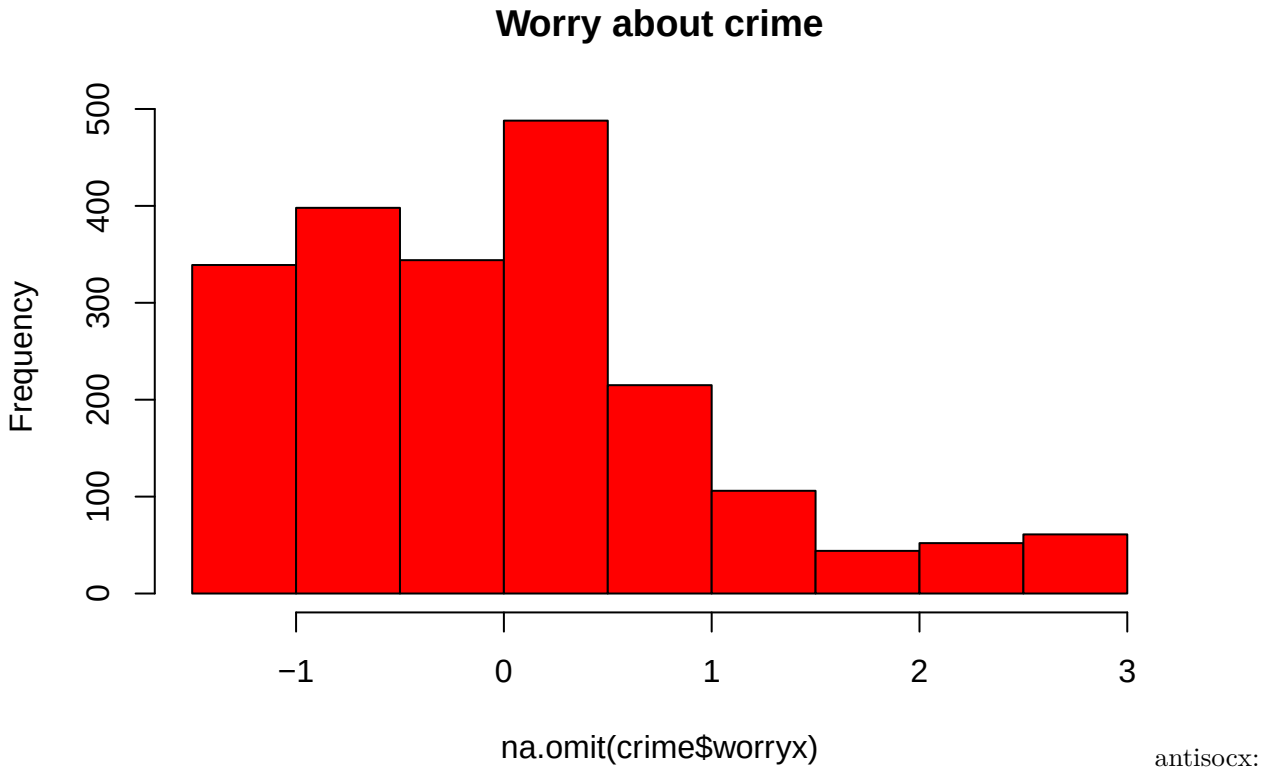
```
##
##                         Male Female
##   Not a victim of crime 3385   4075
##   Victim of crime        652    731
```

Look at the distribution of the two continuous scores

Note: the na.omit command is used to exclude missing data from the computation.

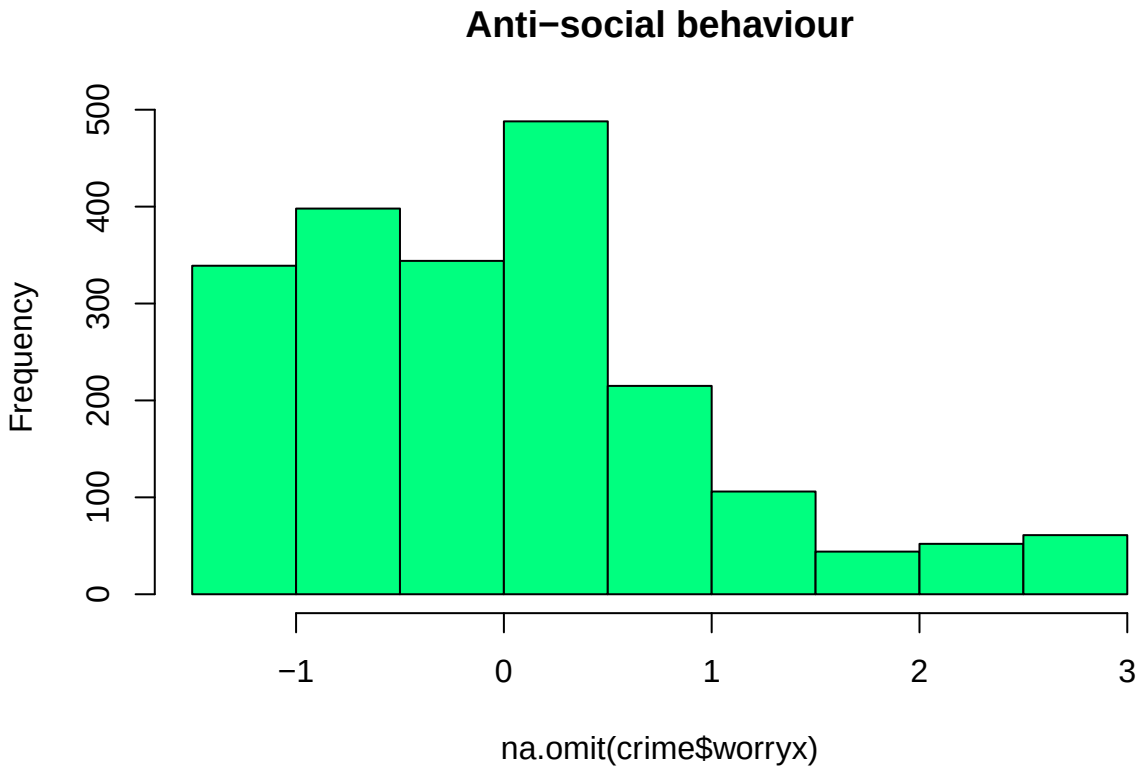
worryx: Worry about being a victim of crime (high score = high level of worry) (Module C)

```
hist(na.omit(crime$worryx),
     main = "Worry about crime",
     col = "red")
```



antisocx: Anti-social behaviour in their neighbourhood (high score =high levels of anti-social behaviour)

```
hist(na.omit(crime$worryx),
     main = "Anti-social behaviour",
     col = "springgreen")
```



Analyze missing data

There are a lot of missing data!!!!

Count missings

```
colSums(is.na(crime)) # Look at the number of missings in each variable
```

```
##  rowlabel     split       sex   yrsarea  resyrago     work2   tenure1  livharm1
##         0         0         0         0      7334         0         0         0
##   agegrp7  ethgrp2a   educat3    rural2 edeprivex wdeprivex IndivWgtx   cause2m
##         0        10        21         0       703      8140         0      6769
##  walkdark   walkday  homealon    wburgl   wmugged  wcarstol  wfromcar    wraped
##      6769      6769      6769      6649      6649      7080      7110      6649
##   wattack  wraceatt    worryx bcsvictim  rubbcomm  vandcomm   poorhou  antisocx
##      6649      6649      6796         0         0         0         0      6694
```

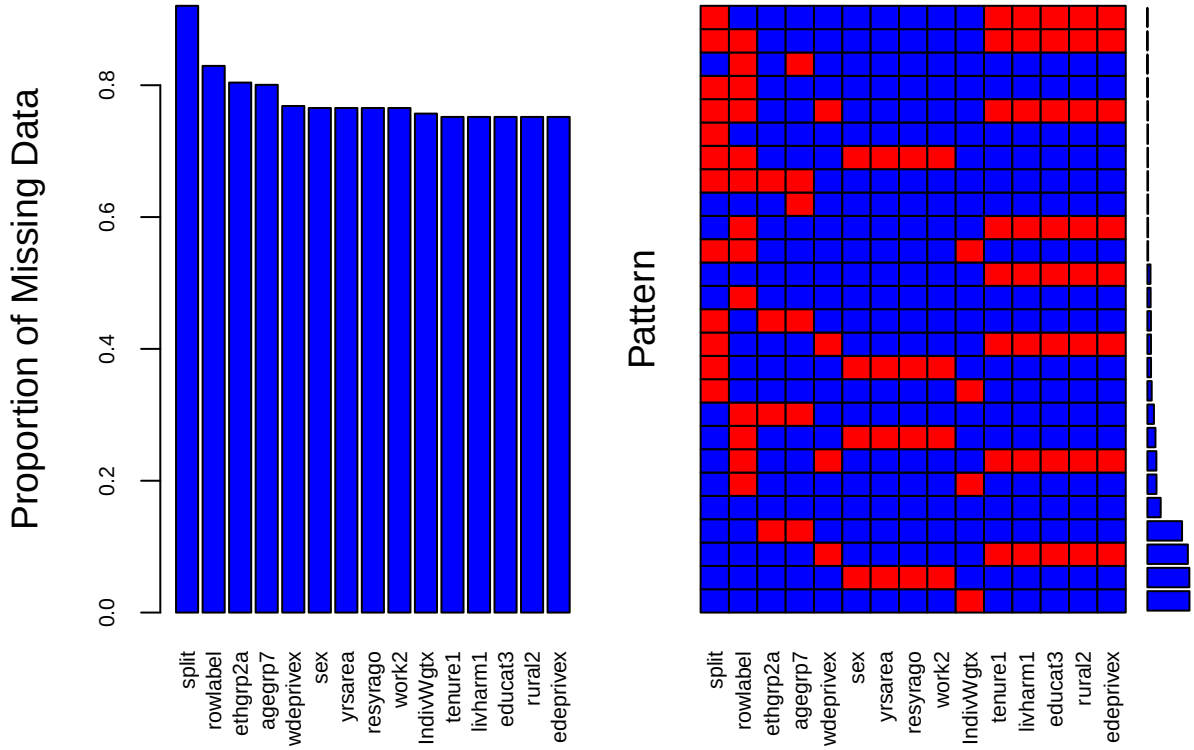
```
table(complete.cases(crime)) # Look atthe number of individuals who have no missings
```

```
##
## FALSE
##  8843
```

Look at patterns of missingness

It allows you to inspect which combinations of variables are likely to be jointly missing

```
MP_plot_crimedata <- aggr(crime[, which(colSums(is.na(crime)) > 1000)],
                          col=c('red','blue'),
                          numbers=TRUE, sortVars=TRUE,
                          labels=names(crime), cex.axis=.7,
                          gap=3, ylab=c("Proportion of Missing Data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##    Variable     Count
##       split 0.9205021
##    rowlabel 0.8293566
##    ethgrp2a 0.8040258
##     agegrp7 0.8006333
##   wdeprivex 0.7685175
##         sex 0.7654642
##     yrsarea 0.7654642
##    resyrago 0.7654642
##       work2 0.7654642
##   IndivWgtx 0.7569829
##     tenure1 0.7518942
##    livharm1 0.7518942
##     educat3 0.7518942
##      rural2 0.7518942
##   edeprivex 0.7518942
```