Module 2 assignment - Statistical Analysis Presentation

Guilherme Amorim

2024-10-17

Initial setup

Load libraries

```
library(haven) # load .sav files

library(tidyverse) # data manipulation and visualisation

library(magrittr) # pipe operator for data manipulation

library(patchwork) # merging plots

library(DescTools) # pre-built function to calculate mode

library(gmodels) # contigency tables

library(flextable) # rendering tables
```

Load data

```
data<-read_sav("Dataset/HSE 2011.sav")
```

Data exploration

General dataset features

```
colnames(data) #  variable names
```

```
##  [1] "hserial"  "pserial"  "HHSize"   "tenureb"  "Sex"      "Age"
##  [7] "MonthAge" "WeekAge"  "PersNo"   "topqual3" "HRPID"    "econact"
## [13] "nssec8"   "Origin"   "totinc"   "eqvinc"   "NurOutc"  "relto01"
## [19] "relto02"  "relto03"  "relto04"  "relto05"  "relto06"  "relto07"
## [25] "relto08"  "relto09"  "Relto10"  "Relto11"  "Relto12"  "ReltoHRP"
## [31] "marstatc" "SHA"      "gor1"     "wt_int"   "wt_nurse" "SayWgt"
```

```
## [37] "SayDiet"  "htval"    "wtval"    "bmival"   "whval"    "omdiaval"
## [43] "omsysval" "dnnow"    "totalwu"  "porfv"    "acutill"  "IllsM1"
## [49] "IllsM2"   "IllsM3"   "IllsM4"   "IllsM5"   "IllsM6"   "limitill"
## [55] "medcnj"   "genhelf2" "cigst1"   "cigst2"
```

```
dim(data) # number rows and columns
```

```
## [1] 10617    58
```

```
str(data) # general features of each variable
```

```
## tibble [10,617 x 58] (S3: tbl_df/tbl/data.frame)
##  $ hserial : num [1:10617] 1e+06 1e+06 1e+06 1e+06 1e+06 ...
##   ..- attr(*, "label")= chr "Serial number of household"
##   ..- attr(*, "format.spss")= chr "F7.0"
##   ..- attr(*, "display_width")= int 9
##  $ pserial : num [1:10617] 1e+08 1e+08 1e+08 1e+08 1e+08 ...
##   ..- attr(*, "label")= chr "Serial number of Individual"
##   ..- attr(*, "format.spss")= chr "F9.0"
##   ..- attr(*, "display_width")= int 11
##  $ HHSize  : num [1:10617] 1 3 2 2 1 2 1 3 3 2 ...
##   ..- attr(*, "label")= chr "(D) Household size"
##   ..- attr(*, "format.spss")= chr "F2.0"
##  $ tenureb : dbl+lbl [1:10617] 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 1, 1, 1, 1...
##    ..@ label        : chr "Household tenure"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:10] -9 -8 -2 -1 1 2 3 4 5 6
##    .. ..- attr(*, "names")= chr [1:10] "Refusal" "Don't Know" "Schedule not applicable" "Item not ap
##  $ Sex     : dbl+lbl [1:10617] 2, 2, 1, 2, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 2, 1, 2, 1...
##    ..@ label        : chr "Sex"
##    ..@ format.spss  : chr "F1.0"
##    ..@ display_width: int 5
##    ..@ labels       : Named num [1:6] -9 -8 -2 -1 1 2
##    .. ..- attr(*, "names")= chr [1:6] "Refusal" "Don't Know" "Schedule not applicable" "Item not appl
##  $ Age     : num [1:10617] 75 47 77 66 44 66 84 63 62 74 ...
##   ..- attr(*, "label")= chr "Age last birthday"
##   ..- attr(*, "format.spss")= chr "F3.0"
##   ..- attr(*, "display_width")= int 5
##  $ MonthAge: num [1:10617] 12 12 12 12 12 12 12 12 12 12 ...
##   ..- attr(*, "label")= chr "Age in months for infants under 1"
##   ..- attr(*, "format.spss")= chr "F2.0"
##   ..- attr(*, "display_width")= int 10
##  $ WeekAge : dbl+lbl [1:10617] 997, 997, 997, 997, 997, 997, 997, 997, 997, 997, 99...
##    ..@ label        : chr "Age in weeks for infants under 2 years"
##    ..@ format.spss  : chr "F3.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num 997
##    .. ..- attr(*, "names")= chr "Over 2 years old"
##  $ PersNo  : num [1:10617] 1 1 1 2 1 1 1 1 2 1 ...
##   ..- attr(*, "label")= chr "Person number"
##   ..- attr(*, "format.spss")= chr "F2.0"
##  $ topqual3: dbl+lbl [1:10617]  6,  4,  1,  1,  3,  1,  7,  7,  4,  2,  4,  4, NA, ...
```

```
##    ..@ label        : chr "(D) Highest Educational Qualification"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 10
##    ..@ labels       : Named num [1:13] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:13] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ HRPID   : dbl+lbl [1:10617] 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 2, 2, 2, 1, 2, 2, 1...
##    ..@ label        : chr "Household Reference Person identifier"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 7
##    ..@ labels       : Named num [1:6] -9 -8 -2 -1 1 2
##    .. ..- attr(*, "names")= chr [1:6] "Refusal" "Don't Know" "Schedule not applicable" "Item not appl
##  $ econact : dbl+lbl [1:10617]  3,  1,  3,  3,  1,  1,  3,  3,  3,  1,  1,  1, NA, ...
##    ..@ label        : chr "(D) Economic Status (4 groups)"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:10] -9 -8 -7 -6 -2 -1 1 2 3 4
##    .. ..- attr(*, "names")= chr [1:10] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ nssec8  : dbl+lbl [1:10617]  6,  1,  1,  2,  2,  1,  4,  3,  3,  3,  6,  7, NA, ...
##    ..@ label      : chr "(D) NS-SEC 8 variable classification (individual)"
##    ..@ format.spss: chr "F3.0"
##    ..@ labels     : Named num [1:15] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:15] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ Origin  : dbl+lbl [1:10617] 1, 1, 1, 1, 1, 9, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
##    ..@ label      : chr "Ethnic origin of individual"
##    ..@ format.spss: chr "F3.0"
##    ..@ labels     : Named num [1:22] -9 -8 -2 -1 1 2 3 4 5 6 ...
##    .. ..- attr(*, "names")= chr [1:22] "Refusal" "Don't Know" "Schedule not applicable" "Item not app
##  $ totinc  : dbl+lbl [1:10617]  6, 97, 97, 97, 97, 97, 96, 96, 96, 97, 96, 96, 96, ...
##    ..@ label      : chr "(D) Total Household Income"
##    ..@ format.spss: chr "F3.0"
##    ..@ labels     : Named num [1:39] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:39] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ eqvinc  : dbl+lbl [1:10617]  10656,     NA,     NA,     NA,     NA,     NA,     ...
##    ..@ label        : chr "(D) Equivalised Income"
##    ..@ format.spss  : chr "F10.2"
##    ..@ display_width: int 12
##    ..@ labels       : Named num [1:2] -90 -1
##    .. ..- attr(*, "names")= chr [1:2] "Age of household member refused" "Item not applicable"
##  $ NurOutc : dbl+lbl [1:10617] NA, 83, 83, 80, 83, NA, 81, NA, NA, NA, 83, 83, 83, ...
##    ..@ label        : chr "Outcome of nurse visit"
##    ..@ format.spss  : chr "F3.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:6] -9 -8 -7 -6 -2 -1
##    .. ..- attr(*, "names")= chr [1:6] "Refused" "Don't know" "Refused/not obtained" "Schedule not obt
##  $ relto01 : dbl+lbl [1:10617] 96, 96, 96,  1, 96, 96, 96, 96,  1, 96, 96,  1,  3, ...
##    ..@ label        : chr "Relationship to person 1"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:29] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:29] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ relto02 : dbl+lbl [1:10617] NA,  8,  1, 96, NA,  1, NA,  1, 96,  1,  1, 96,  3, ...
##    ..@ label        : chr "Relationship to person 2"
##    ..@ format.spss  : chr "F3.0"
##    ..@ display_width: int 9
```

```
##    ..@ labels       : Named num [1:29] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:29] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ relto03 : dbl+lbl [1:10617] NA,  8, NA, NA, NA, NA, NA,  7,  3, NA,  8,  8, 96, ...
##    ..@ label        : chr "Relationship to person 3"
##    ..@ format.spss  : chr "F3.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:29] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:29] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ relto04 : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,  8,  8, 13, ...
##    ..@ label        : chr "Relationship to person 4"
##    ..@ format.spss  : chr "F3.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:29] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:29] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ relto05 : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
##    ..@ label        : chr "Relationship to person 5"
##    ..@ format.spss  : chr "F3.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:29] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:29] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ relto06 : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
##    ..@ label        : chr "Relationship to person 6"
##    ..@ format.spss  : chr "F3.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:29] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:29] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ relto07 : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
##    ..@ label        : chr "Relationship to person 7"
##    ..@ format.spss  : chr "F3.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:29] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:29] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ relto08 : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
##    ..@ label        : chr "Relationship to person 8"
##    ..@ format.spss  : chr "F3.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:29] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:29] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ relto09 : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
##    ..@ label        : chr "Relationship to person 9"
##    ..@ format.spss  : chr "F3.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:29] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:29] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ Relto10 : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
##    ..@ label        : chr "Relationship to person 10"
##    ..@ format.spss  : chr "F3.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:29] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:29] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ Relto11 : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
##    ..@ label        : chr "Relationship to person 11"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 9
```

```
##    ..@ labels       : Named num [1:29] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:29] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ Relto12 : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
##    ..@ label        : chr "Relationship to person 12"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:29] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:29] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ ReltoHRP: dbl+lbl [1:10617] 96, 96, 96,  1, 96, 96, 96,  1, 96, 96, 96,  1,  3, ...
##    ..@ label        : chr "Relationship to Household Reference Person"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 10
##    ..@ labels       : Named num [1:27] -9 -8 -2 -1 1 2 3 4 5 6 ...
##    .. ..- attr(*, "names")= chr [1:27] "Refusal" "Don't Know" "Schedule not applicable" "Item not app
##  $ marstatc: dbl+lbl [1:10617]  5,  5,  2,  2,  1,  2,  6,  2,  2,  2,  2,  2, NA, ...
##    ..@ label        : chr "(D) Marital status including cohabitees"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 10
##    ..@ labels       : Named num [1:13] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:13] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ SHA     : chr [1:10617] "E18000006" "E18000006" "E18000006" "E18000006" ...
##   ..- attr(*, "label")= chr "Strategic Health Authority"
##   ..- attr(*, "format.spss")= chr "A9"
##   ..- attr(*, "display_width")= int 11
##  $ gor1    : dbl+lbl [1:10617] 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6...
##    ..@ label        : chr "Government Office Region - numeric"
##    ..@ format.spss  : chr "F1.0"
##    ..@ display_width: int 6
##    ..@ labels       : Named num [1:11] 1 2 3 4 5 6 7 8 9 10 ...
##    .. ..- attr(*, "names")= chr [1:11] "North East" "North West" "Yorkshire and The Humber" "East Mi
##  $ wt_int  : num [1:10617] 0.886 1.292 0.711 0.707 1.045 ...
##   ..- attr(*, "label")= chr "HSE 2011 Weight for analysis of core interview sample"
##   ..- attr(*, "format.spss")= chr "F4.2"
##  $ wt_nurse: num [1:10617] 0 0 0 0 0 ...
##   ..- attr(*, "label")= chr "hse 2011 Weight for analysis of core nurse sample"
##   ..- attr(*, "format.spss")= chr "F4.2"
##   ..- attr(*, "display_width")= int 10
##  $ SayWgt  : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,  1, ...
##    ..@ label      : chr "How views own weight"
##    ..@ format.spss: chr "F2.0"
##    ..@ labels     : Named num [1:7] -9 -8 -2 -1 1 2 3
##    .. ..- attr(*, "names")= chr [1:7] "Refusal" "Don't Know" "Schedule not applicable" "Item not appl
##  $ SayDiet : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,  3, ...
##    ..@ label        : chr "Whether trying to lose or gain weight"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:7] -9 -8 -2 -1 1 2 3
##    .. ..- attr(*, "names")= chr [1:7] "Refusal" "Don't Know" "Schedule not applicable" "Item not appl
##  $ htval   : dbl+lbl [1:10617] 162,  NA, 170,  NA, 168,  NA,  NA,  NA,  NA,  NA, 16...
##    ..@ label        : chr "(D) Valid height (cm)"
##    ..@ format.spss  : chr "F7.2"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:6] -9 -8 -7 -6 -2 -1
##    .. ..- attr(*, "names")= chr [1:6] "Refused" "Don't know" "Refused/not obtained" "Schedule not obt
```

```
##  $ wtval   : dbl+lbl [1:10617]  66.3,    NA,  74.2,    NA,    NA,    NA,    NA,    ...
##    ..@ label        : chr "(D) Valid weight (Kg) inc. estimated>130kg"
##    ..@ format.spss  : chr "F7.2"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:6] -9 -8 -7 -6 -2 -1
##    .. ..- attr(*, "names")= chr [1:6] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ bmival  : dbl+lbl [1:10617] 25.3,   NA, 25.6,   NA,   NA,   NA,   NA,   NA,   NA...
##    ..@ label      : chr "(D) Valid BMI"
##    ..@ format.spss: chr "F6.2"
##    ..@ labels     : Named num [1:6] -9 -8 -7 -6 -2 -1
##    .. ..- attr(*, "names")= chr [1:6] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ whval   : dbl+lbl [1:10617]    NA,    NA,    NA,    NA,    NA,    NA, 0.938,    ...
##    ..@ label        : chr "(D) Valid Mean Waist/Hip ratio"
##    ..@ format.spss  : chr "F5.2"
##    ..@ display_width: int 7
##    ..@ labels       : Named num [1:6] -9 -8 -7 -6 -2 -1
##    .. ..- attr(*, "names")= chr [1:6] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ omdiaval: dbl+lbl [1:10617]   NA,   NA,   NA,   NA,   NA,   NA, 90.0,   NA,   NA...
##    ..@ label        : chr "(D) Omron Valid Mean Diastolic BP"
##    ..@ format.spss  : chr "F7.2"
##    ..@ display_width: int 10
##    ..@ labels       : Named num [1:6] -9 -8 -7 -6 -2 -1
##    .. ..- attr(*, "names")= chr [1:6] "Refused" "Don't know" "Refused, attempted but not obtained, n
##  $ omsysval: dbl+lbl [1:10617]  NA,  NA,  NA,  NA,  NA,  NA, 163,  NA,  NA,  NA,  N...
##    ..@ label        : chr "(D) Omron Valid Mean Systolic BP"
##    ..@ format.spss  : chr "F7.2"
##    ..@ display_width: int 10
##    ..@ labels       : Named num [1:6] -9 -8 -7 -6 -2 -1
##    .. ..- attr(*, "names")= chr [1:6] "Refused" "Don't know" "Refused, attempted but not obtained, n
##  $ dnnow   : dbl+lbl [1:10617]  2,  1,  1,  2,  1,  1,  1,  2,  1,  1,  1,  1, NA, ...
##    ..@ label        : chr "Whether drink nowadays"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 7
##    ..@ labels       : Named num [1:5] -9 -8 -1 1 2
##    .. ..- attr(*, "names")= chr [1:5] "Refusal" "Don't know" "Item not applicable" "Yes" ...
##  $ totalwu : dbl+lbl [1:10617]  0.058,  4.991, 49.029,  0.000, 30.230, 13.558, 24.6...
##    ..@ label        : chr "(D) Total units of alcohol/week"
##    ..@ format.spss  : chr "F7.2"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:3] -9 -8 -1
##    .. ..- attr(*, "names")= chr [1:3] "Refused/not answered" "Don't know" "Item not applicable"
##  $ porfv   : dbl+lbl [1:10617]  4.00,  6.50,  1.00,  2.00, 10.33,  5.33,  5.00,  2....
##    ..@ label      : chr "(D) Total portion of fruit and veg"
##    ..@ format.spss: chr "F6.2"
##    ..@ labels     : Named num [1:6] -9 -8 -7 -6 -2 -1
##    .. ..- attr(*, "names")= chr [1:6] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ acutill : dbl+lbl [1:10617] 1, 1, 1, 5, 1, 1, 1, 1, 1, 1, 5, 1, 1, 1, 3, 1, 1, 1...
##    ..@ label        : chr "(D) Acute sickness last two weeks"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 9
##    ..@ labels       : Named num [1:11] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
##    .. ..- attr(*, "names")= chr [1:11] "Refused" "Don't know" "Refused/not obtained" "Schedule not o
##  $ IllsM1  : dbl+lbl [1:10617] 34, NA, NA,  8, NA, NA, NA, NA, NA, 17, 36, NA, NA, ...
##    ..@ label      : chr "Type of illness - 1st"
```

```
##    ..@ format.spss: chr "F3.0"
##    ..@ labels     : Named num [1:45] -1 1 2 3 4 5 6 7 8 9 ...
##    .. ..- attr(*, "names")= chr [1:45] "Item not applicable" "Cancer (neoplasm)" "Diabetes" "Other er
##  $ IllsM2  : dbl+lbl [1:10617] 97, NA, NA, 97, NA, NA, NA, NA, NA, 18, 97, NA, NA, ...
##    ..@ label      : chr "Type of illness - 2nd"
##    ..@ format.spss: chr "F3.0"
##    ..@ labels     : Named num [1:45] -1 1 2 3 4 5 6 7 8 9 ...
##    .. ..- attr(*, "names")= chr [1:45] "Item not applicable" "Cancer (neoplasm)" "Diabetes" "Other er
##  $ IllsM3  : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, 97, NA, NA, NA, ...
##    ..@ label      : chr "Type of illness - 3rd"
##    ..@ format.spss: chr "F3.0"
##    ..@ labels     : Named num [1:45] -1 1 2 3 4 5 6 7 8 9 ...
##    .. ..- attr(*, "names")= chr [1:45] "Item not applicable" "Cancer (neoplasm)" "Diabetes" "Other er
##  $ IllsM4  : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
##    ..@ label      : chr "Type of illness - 4th"
##    ..@ format.spss: chr "F3.0"
##    ..@ labels     : Named num [1:45] -1 1 2 3 4 5 6 7 8 9 ...
##    .. ..- attr(*, "names")= chr [1:45] "Item not applicable" "Cancer (neoplasm)" "Diabetes" "Other er
##  $ IllsM5  : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
##    ..@ label      : chr "Type of illness - 5th"
##    ..@ format.spss: chr "F3.0"
##    ..@ labels     : Named num [1:45] -1 1 2 3 4 5 6 7 8 9 ...
##    .. ..- attr(*, "names")= chr [1:45] "Item not applicable" "Cancer (neoplasm)" "Diabetes" "Other er
##  $ IllsM6  : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
##    ..@ label      : chr "Type of illness - 6th"
##    ..@ format.spss: chr "F3.0"
##    ..@ labels     : Named num [1:45] -1 1 2 3 4 5 6 7 8 9 ...
##    .. ..- attr(*, "names")= chr [1:45] "Item not applicable" "Cancer (neoplasm)" "Diabetes" "Other er
##  $ limitill: dbl+lbl [1:10617] 2, 3, 3, 1, 3, 3, 3, 3, 3, 2, 1, 3, 3, 3, 1, 2, 3, 2...
##    ..@ label        : chr "(D) Limiting longstanding illness"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 10
##    ..@ labels       : Named num [1:9] -9 -8 -7 -6 -2 -1 1 2 3
##    .. ..- attr(*, "names")= chr [1:9] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ medcnj  : dbl+lbl [1:10617] NA, NA, NA, NA, NA, NA,  1, NA, NA, NA, NA, NA, NA, ...
##    ..@ label      : chr "(D) Whether taking medication - excluding contraceptives only"
##    ..@ format.spss: chr "F2.0"
##    ..@ labels     : Named num [1:9] -9 -8 -7 -6 -2 -1 1 2 3
##    .. ..- attr(*, "names")= chr [1:9] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ genhelf2: dbl+lbl [1:10617] 1, 1, 1, 3, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 1, 1...
##    ..@ label        : chr "(D) Self-assessed general health - grouped"
##    ..@ format.spss  : chr "F2.0"
##    ..@ display_width: int 10
##    ..@ labels       : Named num [1:4] -8 1 2 3
##    .. ..- attr(*, "names")= chr [1:4] "Dont know" "Very good/good" "Fair" "Bad/very bad"
##  $ cigst1  : dbl+lbl [1:10617]  3,  3,  1,  1,  2,  3,  3,  1,  3,  3,  1,  4, NA, ...
##    ..@ label      : chr "(D) Cigarette Smoking Status - Never/Ex-reg/Ex-occ/Current"
##    ..@ format.spss: chr "F2.0"
##    ..@ labels     : Named num [1:10] -9 -8 -7 -6 -2 -1 1 2 3 4
##    .. ..- attr(*, "names")= chr [1:10] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
##  $ cigst2  : dbl+lbl [1:10617]  5,  5,  5,  5,  5,  5,  5,  5,  5,  5,  5,  1, NA, ...
##    ..@ label      : chr "(D) Cigarette Smoking Status - Banded current smokers"
##    ..@ format.spss: chr "F2.0"
##    ..@ labels     : Named num [1:11] -9 -8 -7 -6 -2 -1 1 2 3 4 ...
```

```
##    .. ..- attr(*, "names")= chr [1:11] "Refused" "Don't know" "Refused/not obtained" "Schedule not ob
```

```
head(data) # first 6 rows
```

```
## # A tibble: 6 x 58
##   hserial   pserial HHSize tenureb       Sex       Age MonthAge WeekAge   PersNo
##     <dbl>     <dbl>  <dbl> <dbl+lbl>     <dbl+l> <dbl>    <dbl> <dbl+lbl>  <dbl>
## 1 1001011 100101101      1 1 [Own it ou~ 2 [Fem~    75       12 997 [Ove~      1
## 2 1001031 100103101      3 1 [Own it ou~ 2 [Fem~    47       12 997 [Ove~      1
## 3 1001041 100104101      2 1 [Own it ou~ 1 [Mal~    77       12 997 [Ove~      1
## 4 1001041 100104102      2 1 [Own it ou~ 2 [Fem~    66       12 997 [Ove~      2
## 5 1001051 100105101      1 1 [Own it ou~ 1 [Mal~    44       12 997 [Ove~      1
## 6 1001061 100106101      2 1 [Own it ou~ 1 [Mal~    66       12 997 [Ove~      1
## # i 49 more variables: topqual3 <dbl+lbl>, HRPID <dbl+lbl>, econact <dbl+lbl>,
## #   nssec8 <dbl+lbl>, Origin <dbl+lbl>, totinc <dbl+lbl>, eqvinc <dbl+lbl>,
## #   NurOutc <dbl+lbl>, relto01 <dbl+lbl>, relto02 <dbl+lbl>, relto03 <dbl+lbl>,
## #   relto04 <dbl+lbl>, relto05 <dbl+lbl>, relto06 <dbl+lbl>, relto07 <dbl+lbl>,
## #   relto08 <dbl+lbl>, relto09 <dbl+lbl>, Relto10 <dbl+lbl>, Relto11 <dbl+lbl>,
## #   Relto12 <dbl+lbl>, ReltoHRP <dbl+lbl>, marstatc <dbl+lbl>, SHA <chr>,
## #   gor1 <dbl+lbl>, wt_int <dbl>, wt_nurse <dbl>, SayWgt <dbl+lbl>, ...
```

The data includes 58 columns and 10617 rows (observations). The main variables of interest for subsequent analyses are:
- HHSize (household size) - Sex - Age (age at last birthday) - topqual3 (Highest Educational Qualification) - totinc (Total Household Income) - marstatc (Marital status including cohabitees) - htval (Valid height (cm)) - wtval (Valid weight (Kg) inc. estimated>130kg) - bmival (Valid BMI) - dnnow (Whether drink nowadays) - totalwu (total units of alcohol/week) - gor1 (Government Office Region - numeric)

Restrict dataset to variables of interest and rename them

```
data%<>%
  select(age=Age,
         sex=Sex,
         household_size=HHSize,
         education=topqual3,
         household_income=totinc,
         marital_status=marstatc,
         height=htval,
         weight=wtval,
         bmi=bmival,
         drinks=dnnow,
         alcohol_units=totalwu,
         region=gor1)
```

Descriptive statistics

Age

```r
noquote("Summary statistics:")
```

```
## [1] Summary statistics:
```

```r
summary(data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   22.00   42.00   41.56   61.00  100.00
```

```r
noquote("Standard deviation:")
```

```
## [1] Standard deviation:
```

```r
sd(data$age)
```

```
## [1] 23.83203
```

```r
noquote(paste("Mode:", Mode(data$age)))
```

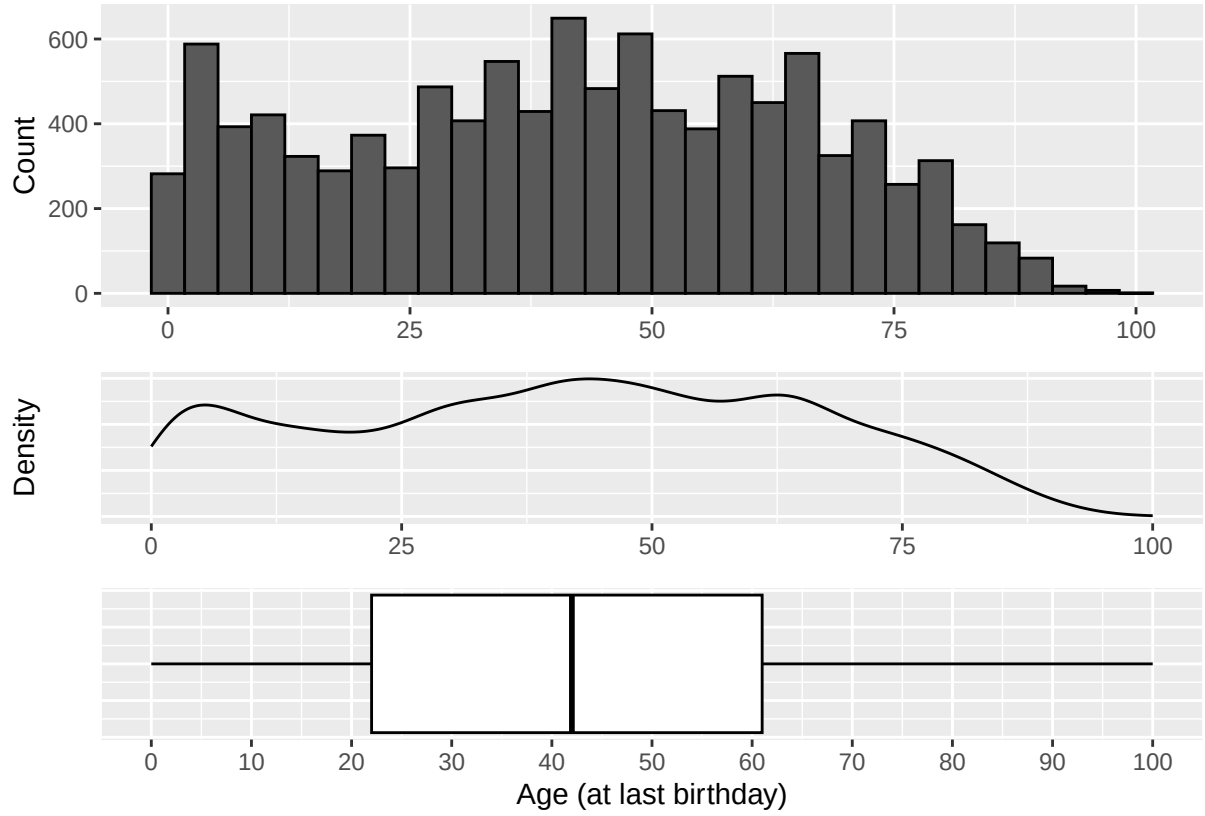
```
## [1] Mode: 42 Mode: 64
```

```r
# calculate mode
noquote(paste0("Mode: ", round(with(density(data$age, na.rm = T), x[which.max(y)]),2))) ##   kernel den
```

```
## [1] Mode: 43.77
```

```r
p1<-data%>%
  ggplot(aes(age))+
  geom_histogram(color="black")+
  theme(axis.title.x = element_blank())+
  labs(y="Count")

p2<-data%>%
  ggplot(aes(age))+
  geom_density(color="black")+
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        axis.title.x = element_blank())+
  labs(y="Density")

p3<-data%>%
  ggplot(aes(age))+
  geom_boxplot(color="black")+
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())+
  scale_x_continuous(breaks=seq(0,100,10))+
  labs(x="Age (at last birthday)")
```

```
p1/p2/p3+plot_layout(nrow = 3, heights = c(2, 1,1))->age_plot

age_plot
```




```
ggsave("Outputs/age_plot.png",
       width = 15,
       height=10,
       units="cm")
```

Sex

```
table(data$sex)
```

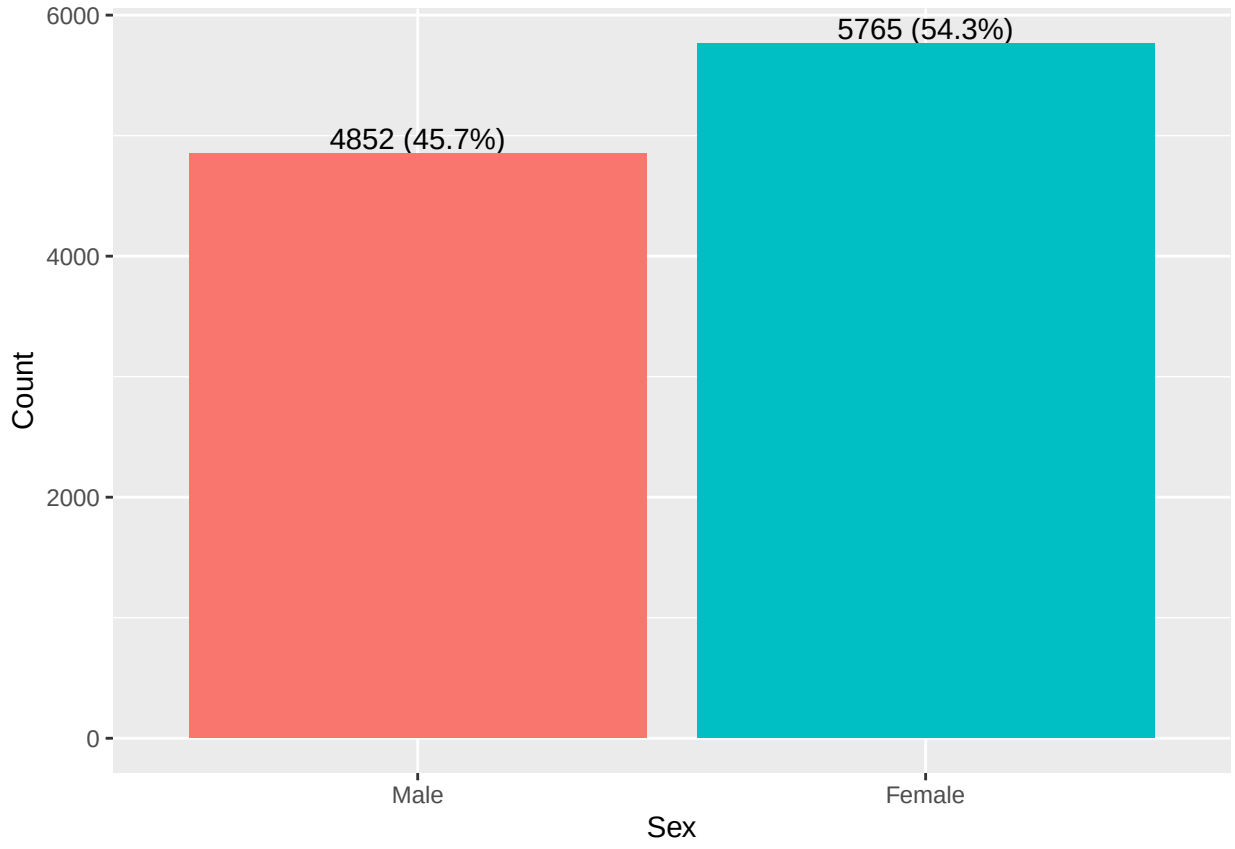
```
##
##    1    2
## 4852 5765
```

```
attr(data$sex, "labels")
```

```
##                 Refusal              Don't Know Schedule not applicable
```

```
##                      -9                      -8                      -2
##     Item not applicable                    Male                  Female
##                      -1                       1                       2
```

```
data%>%
  ggplot(aes(as_factor(sex), fill=as_factor(sex)))+
  geom_text(stat="count",aes(label=paste0(after_stat(count), " (", round(after_stat(count)/length(data$
            vjust=-0.2)+
  geom_bar()+
  labs(x="Sex",
       y="Count")+
  theme(legend.position = "none")->sex_plot

sex_plot
```




```
ggsave("Outputs/sex_plot.png")
```

Marital status

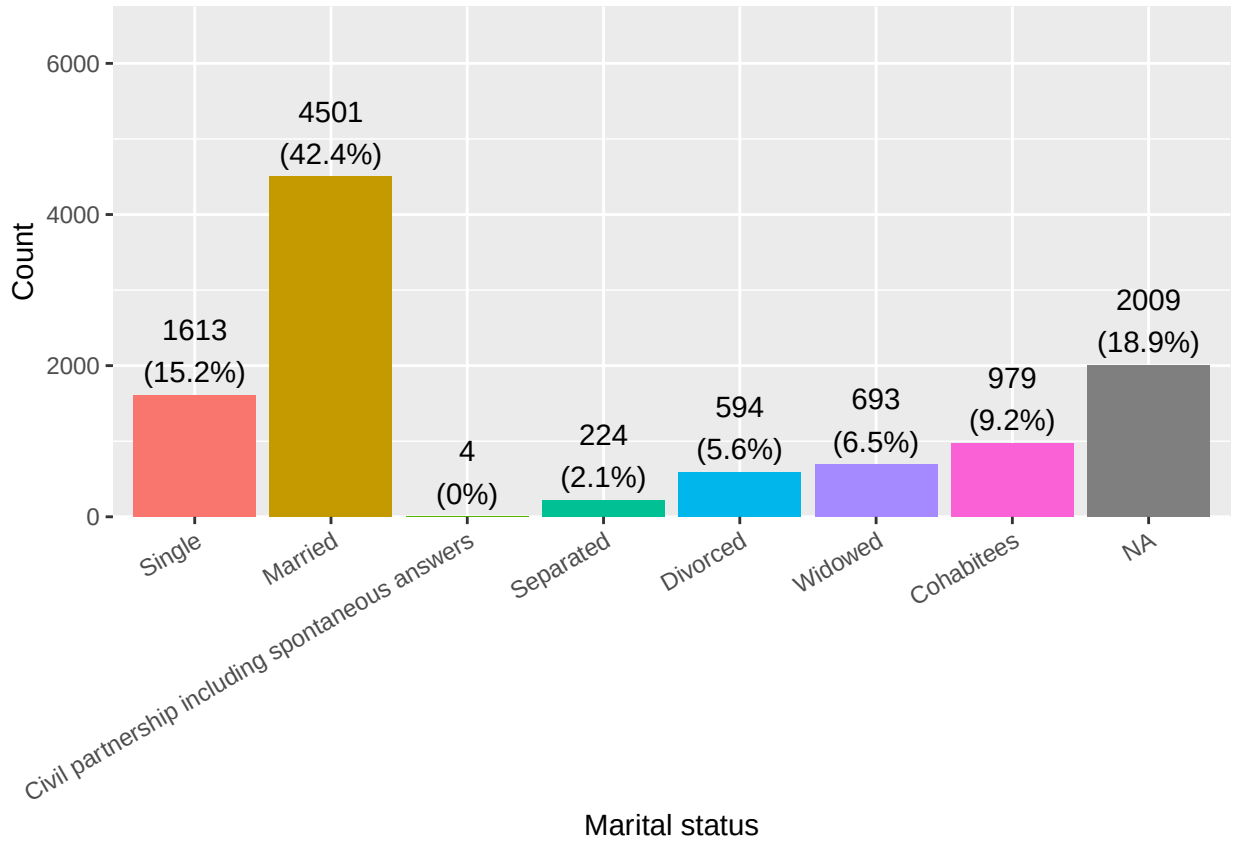
```
table(data$marital_status)
```

```
##
##    1    2    3    4    5    6    7
## 1613 4501    4  224  594  693  979
```

```
attr(data$marital_status, "labels")
```

```
##                                          Refused
##                                               -9
##                                       Don't know
##                                               -8
##                                Refused/not obtained
##                                               -7
##                              Schedule not obtained
##                                               -6
##                            Schedule not applicable
##                                               -2
##                                     Not applicable
##                                               -1
##                                           Single
##                                                1
##                                          Married
##                                                2
## Civil partnership including spontaneous answers
##                                                3
##                                        Separated
##                                                4
##                                         Divorced
##                                                5
##                                          Widowed
##                                                6
##                                       Cohabitees
##                                                7
```

```
data%>%
  ggplot(aes(as_factor(marital_status), fill=as_factor(marital_status)))+
  geom_bar()+
  geom_text(stat="count",aes(label=paste0(after_stat(count), "\n(", round(after_stat(count)/length(data$
            vjust=-0.2)+

  labs(x="Marital status",
       y="Count")+
  theme(axis.text.x = element_text(angle=30, hjust=1,vjust=1),
        legend.position = "none")+
  scale_y_continuous(expand=expansion(c(0,0.5)))->marital_status_plot

marital_status_plot
```




```
ggsave("Outputs/marital_status_plot.png")
```

Education

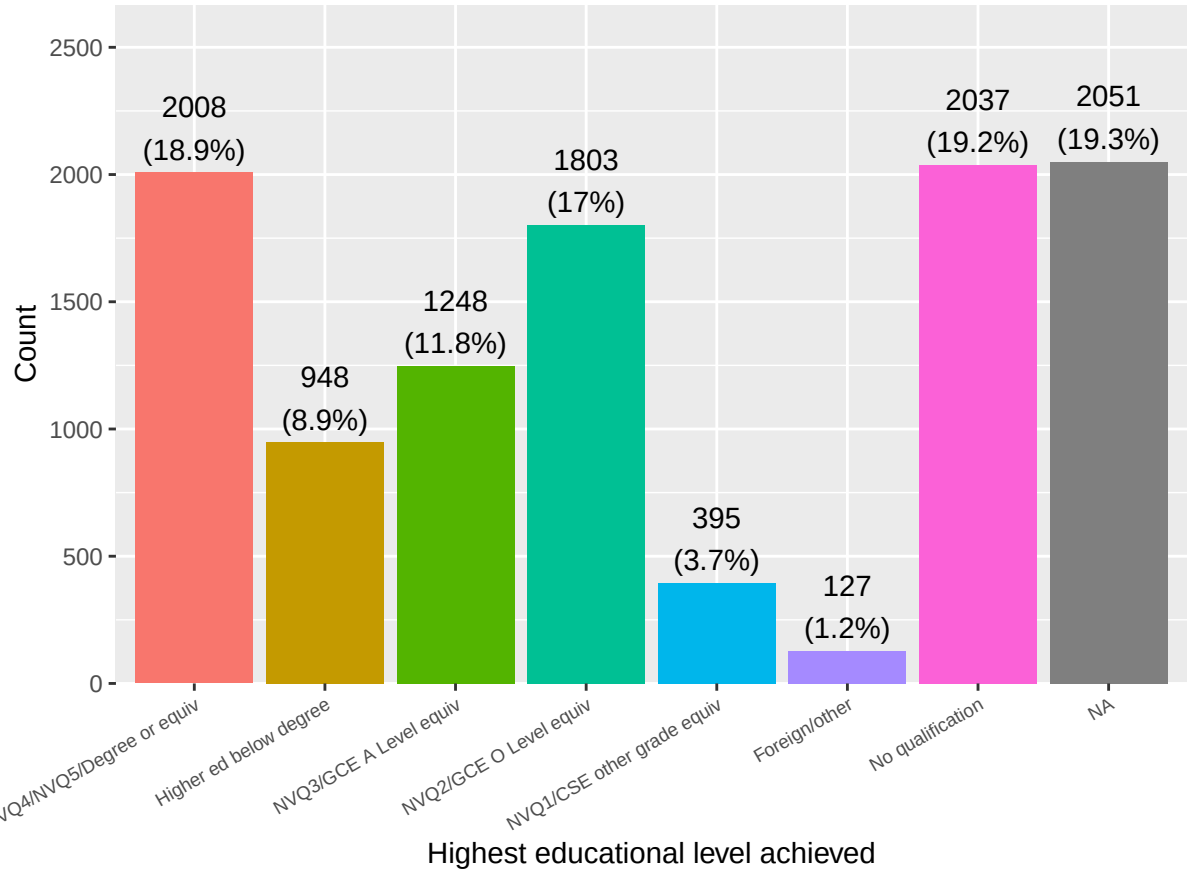
```
table(as_factor(data$education))
```

```
##
##                     Refused                Don't know
##                           0                         0
##       Refused/not obtained      Schedule not obtained
##                           0                         0
##    Schedule not applicable             Not applicable
##                           0                         0
##  NVQ4/NVQ5/Degree or equiv     Higher ed below degree
##                        2008                       948
##     NVQ3/GCE A Level equiv     NVQ2/GCE O Level equiv
##                        1248                      1803
## NVQ1/CSE other grade equiv             Foreign/other
##                         395                       127
##           No qualification
##                        2037
```

```r
attr(data$education, "labels")
```

```
##                     Refused                  Don't know
##                          -9                          -8
##       Refused/not obtained       Schedule not obtained
##                          -7                          -6
##    Schedule not applicable             Not applicable
##                          -2                          -1
##  NVQ4/NVQ5/Degree or equiv     Higher ed below degree
##                           1                           2
##     NVQ3/GCE A Level equiv     NVQ2/GCE O Level equiv
##                           3                           4
## NVQ1/CSE other grade equiv              Foreign/other
##                           5                           6
##           No qualification
##                           7
```

```r
data%>%
  ggplot(aes(as_factor(education), fill=as_factor(education)))+
  geom_bar()+
  geom_text(stat="count",aes(label=paste0(after_stat(count), "\n(", round(after_stat(count)/length(data$
            vjust=-0.2)+

  labs(x="Highest educational level achieved",
       y="Count")+
  theme(axis.text.x = element_text(angle=30, hjust=1,vjust=1, size=7),
        legend.position = "none",
        plot.margin = unit(c(0, 0, 0, 1),"cm"))+
  scale_y_continuous(expand=expansion(c(0,0.3)))->education_plot

education_plot
```




```
ggsave("Outputs/education_plot.png",
       width = 15,
       height=10,
       units = "cm")
```

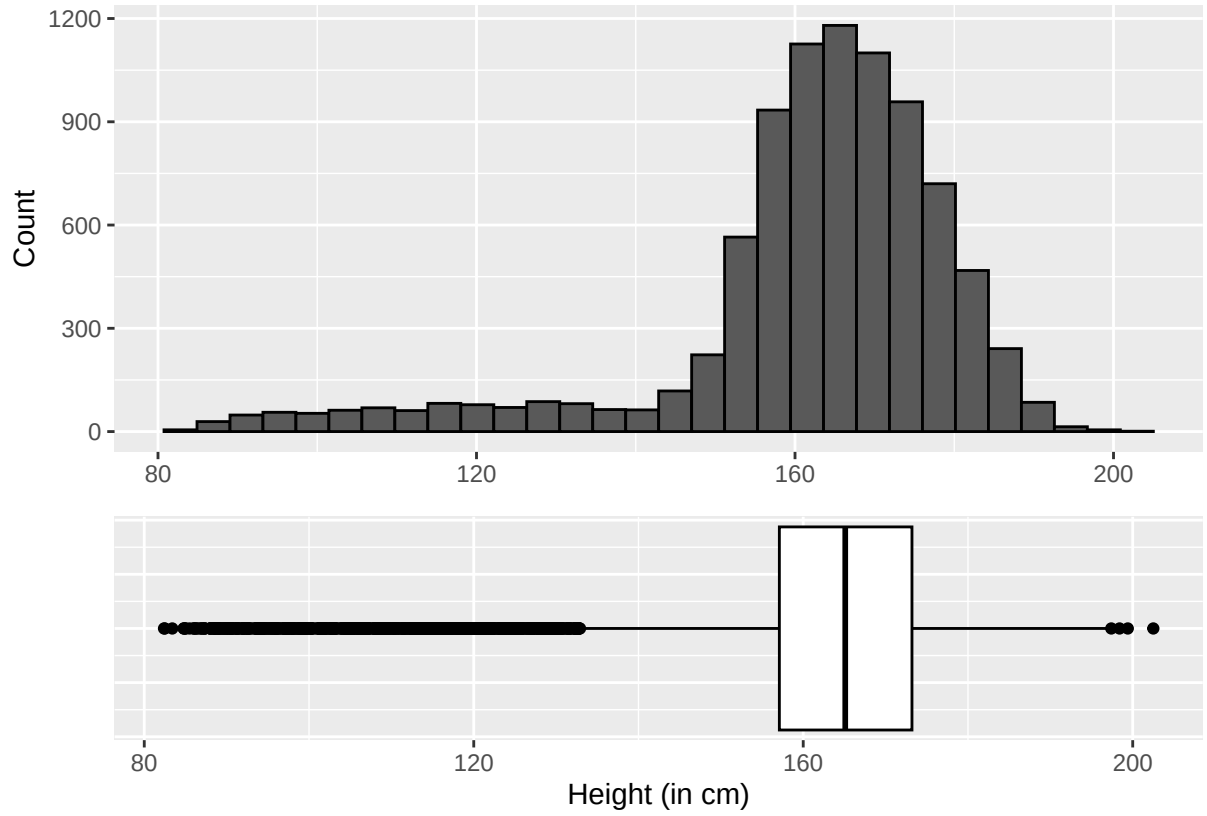
Height

```
summary(data$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    82.4   157.1   165.1   161.8   173.2   202.5    1971
```

```
p1<-data%>%
  ggplot(aes(height))+
  geom_histogram(color="black")+
  theme(axis.title.x = element_blank())+
  labs(y="Count")

p2<-data%>%
  ggplot(aes(height))+
  geom_boxplot(color="black")+
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())+
```

```
  labs(x="Height (in cm)")


p1/p2+plot_layout(nrow = 2, heights = c(2, 1))->height_plot

height_plot
```




```
ggsave("Outputs/height_plot.png")
```

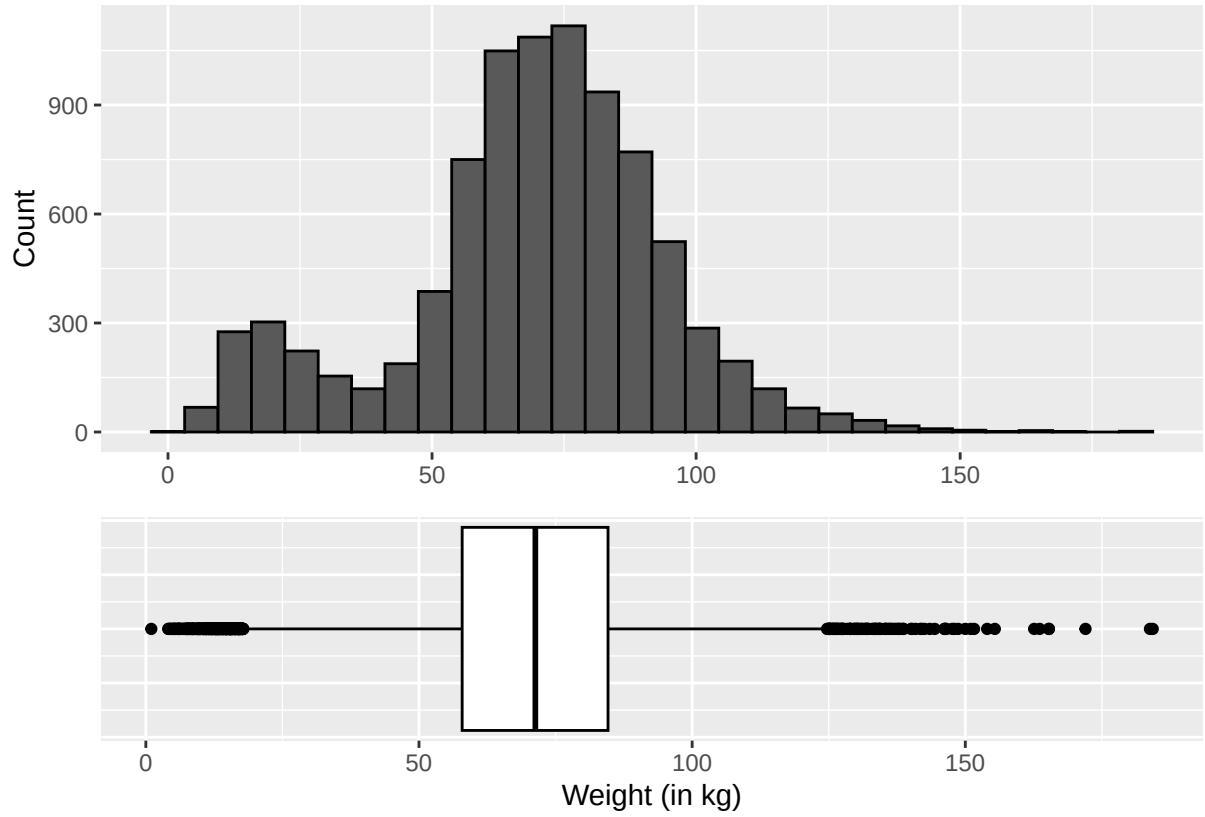
Weight

```
summary(data$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   57.90   71.30   69.15   84.60  184.30    1876
```

```
p1<-data%>%
  ggplot(aes(weight))+
  geom_histogram(color="black")+
  theme(axis.title.x = element_blank())+
  labs(y="Count")
```

```r
p2<-data%>%
  ggplot(aes(weight))+
  geom_boxplot(color="black")+
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())+
  labs(x="Weight (in kg)")


p1/p2+plot_layout(nrow = 2, heights = c(2, 1))->weight_plot

weight_plot
```




```r
ggsave("Outputs/weight_plot.png")
```

BMI

```r
noquote("Summary statistics:")
```

```
## [1] Summary statistics:
```

```r
summary(data$bmi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    8.34   21.93   25.59   25.92   29.39   65.28    2241
```

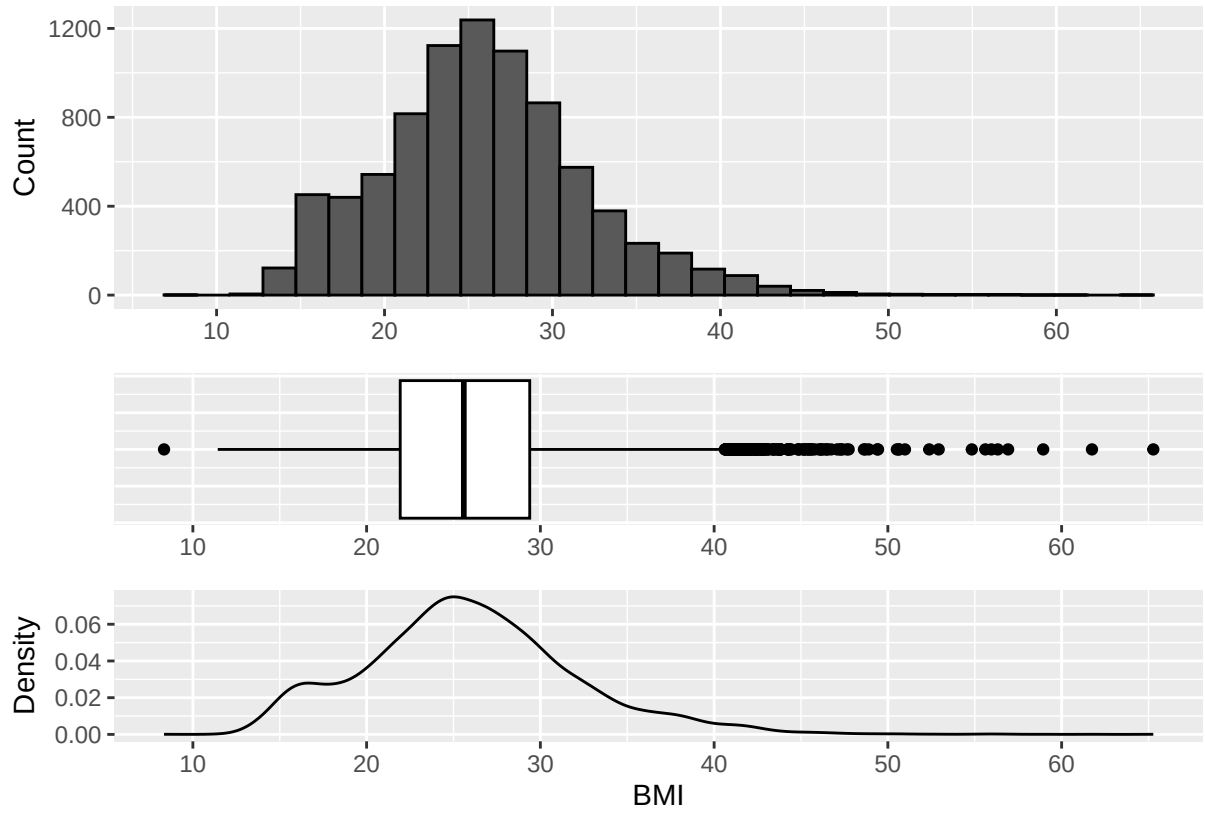
```
noquote(paste("Standard deviation:", sd(data$bmi, na.rm = T)))
```

```
## [1] Standard deviation: 6.13834351764461
```

```
# calculate mode
noquote(paste0("Mode: ", round(with(density(data$bmi, na.rm = T), x[which.max(y)]),2))) ##   kernel den
```

```
## [1] Mode: 25
```

```
p1<-data%>%
  ggplot(aes(bmi))+
  geom_histogram(color="black")+
  theme(axis.title.x = element_blank())+
  labs(y="Count")+
  scale_x_continuous(breaks = seq(10,60,10))

p2<-data%>%
  ggplot(aes(bmi))+
  geom_boxplot(color="black")+
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        axis.title.x = element_blank())+
  scale_x_continuous(breaks = seq(10,60,10))

p3<-data%>%
  ggplot(aes(bmi))+
  geom_density(color="black")+
  scale_x_continuous(breaks = seq(10,60,10))+
  labs(y="Density",
       x="BMI")


p1/p2/p3+plot_layout(nrow = 3, heights = c(2, 1,1))->bmi_plot

bmi_plot
```




```
ggsave("Outputs/bmi_plot.png")
```

Household size

```
table(data$household_size)
```

```
##
##    1    2    3    4    5    6    7    8    9   10
## 1594 3490 2067 2406  719  218   61   37    9   16
```

```
noquote("Summary statistics:")
```

```
## [1] Summary statistics:
```

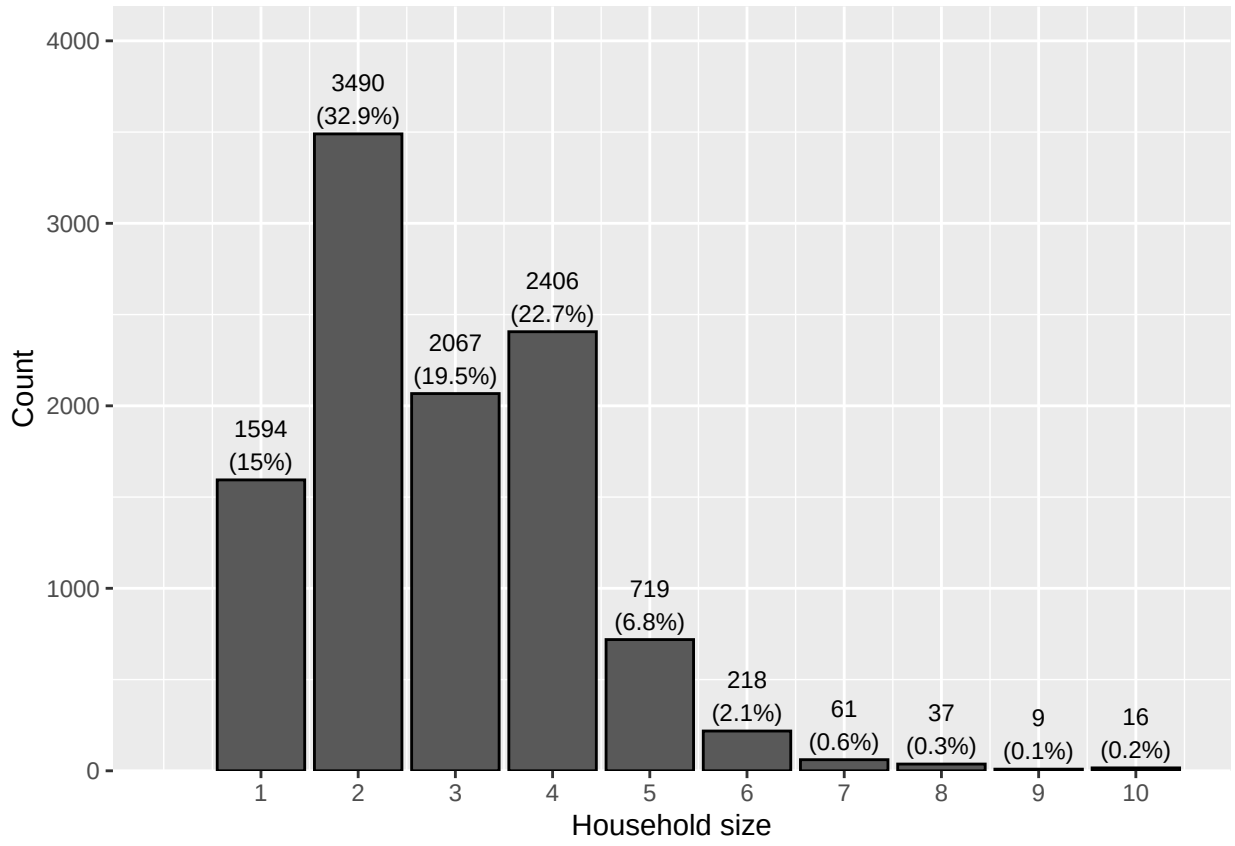
```
summary(data$household_size)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.851   4.000  10.000
```

```
noquote(paste("Standard deviation:", sd(data$household_size, na.rm = T)))
```

```
## [1] Standard deviation: 1.36852803541081
```

```
data%>%
  ggplot(aes(household_size))+
  geom_bar(color="black")+
   geom_text(stat="count",aes(label=paste0(after_stat(count), "\n(", round(after_stat(count)/length(data
             vjust=-0.2,
             size=3)+
  scale_x_continuous(limits=c(0,NA), breaks=seq(1,10,1))+
  labs(x="Household size",
       y="Count")+
    scale_y_continuous(expand=expansion(c(0,0.2)))->household_size_plot

household_size_plot
```



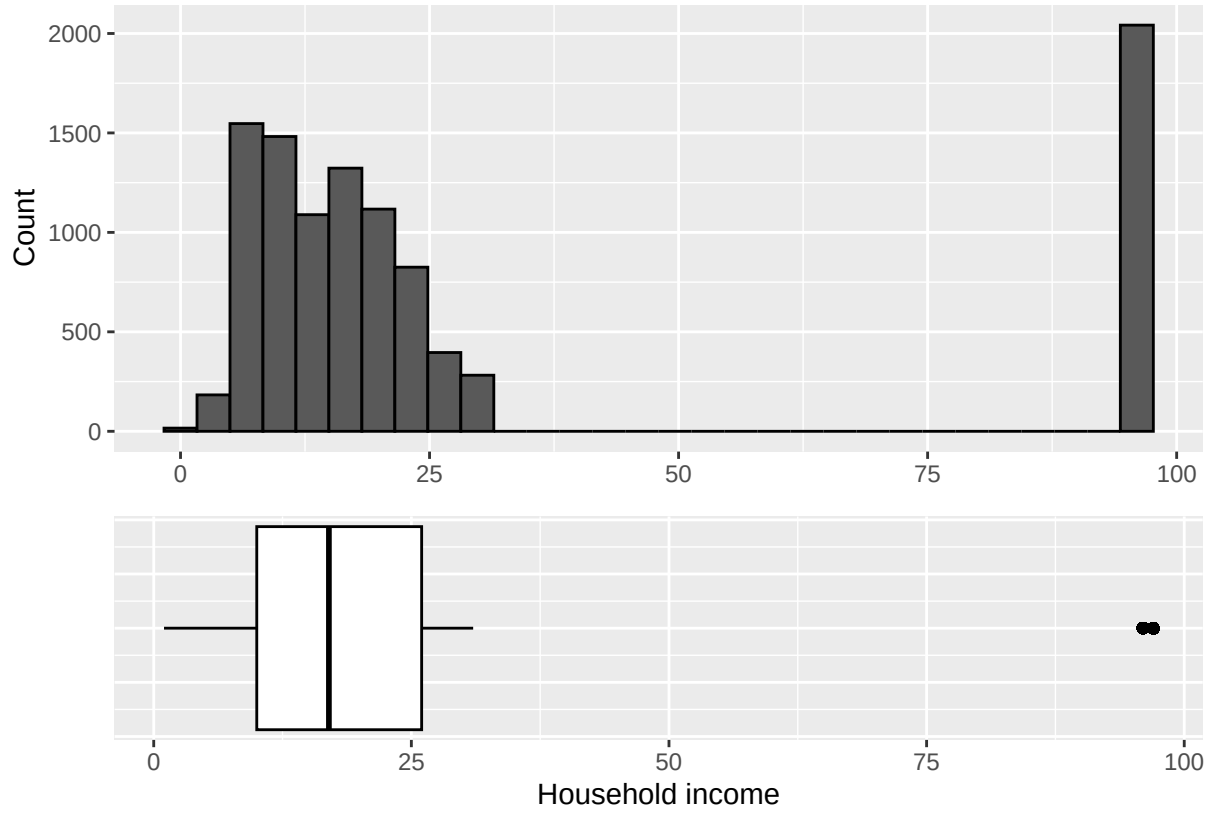

```
ggsave("Outputs/household_size_plot.png")
```

Household income

```
summary(data$household_income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   10.00   17.00   30.97   26.00   97.00     315
```

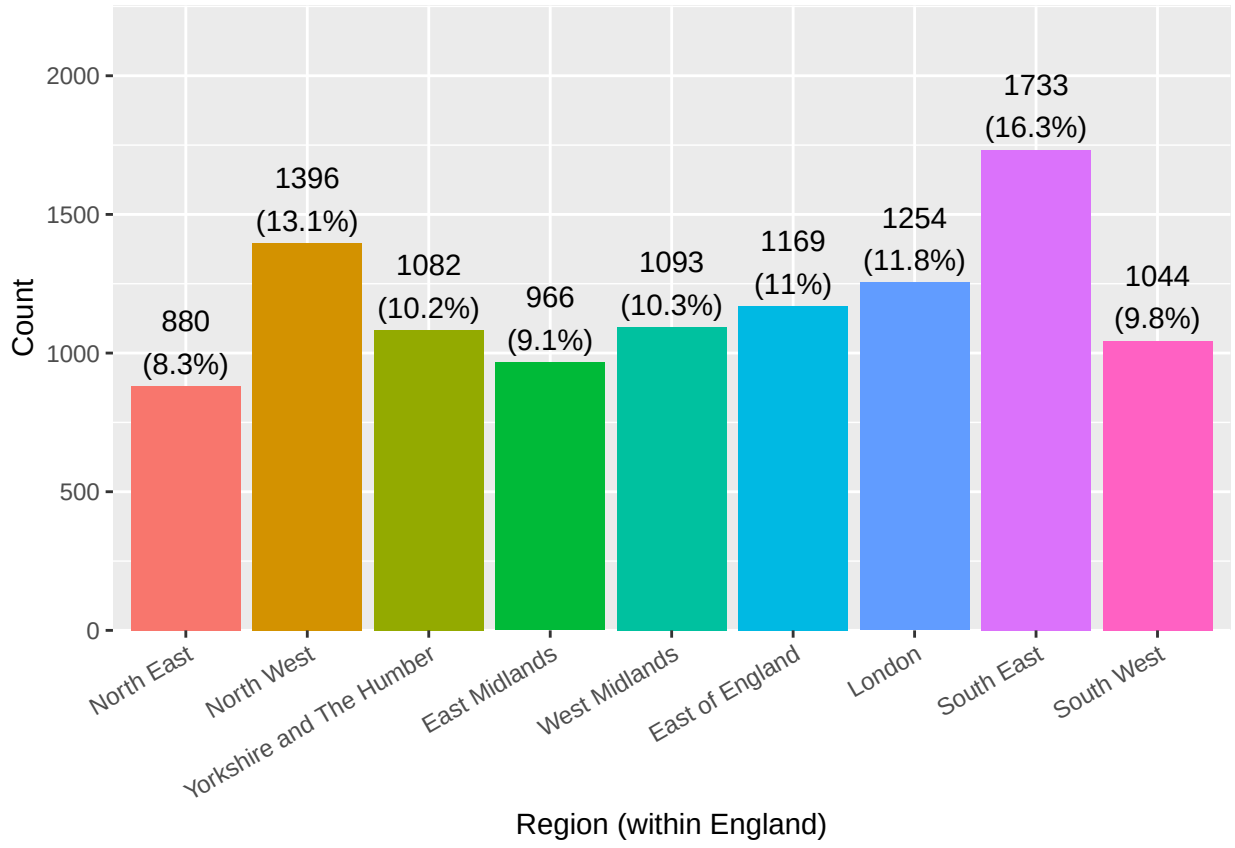
```
p1<-data%>%
  ggplot(aes(household_income))+
  geom_histogram(color="black")+
  theme(axis.title.x = element_blank())+
  labs(y="Count")

p2<-data%>%
  ggplot(aes(household_income))+
  geom_boxplot(color="black")+
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())+
  labs(x="Household income")


p1/p2+plot_layout(nrow = 2, heights = c(2, 1))->household_income_plot

household_income_plot
```




```
ggsave("Outputs/household_income_plot.png")
```

Region

```
data%>%
  ggplot(aes(as_factor(region), fill=as_factor(region)))+
  geom_bar()+
  geom_text(stat="count",aes(label=paste0(after_stat(count), "\n(", round(after_stat(count)/length(data$
            vjust=-0.2)+

  labs(x="Region (within England)",
       y="Count")+
  theme(axis.text.x = element_text(angle=30, hjust=1,vjust=1),
        legend.position = "none")+
  scale_y_continuous(expand=expansion(c(0,0.3)))->region_plot

region_plot
```

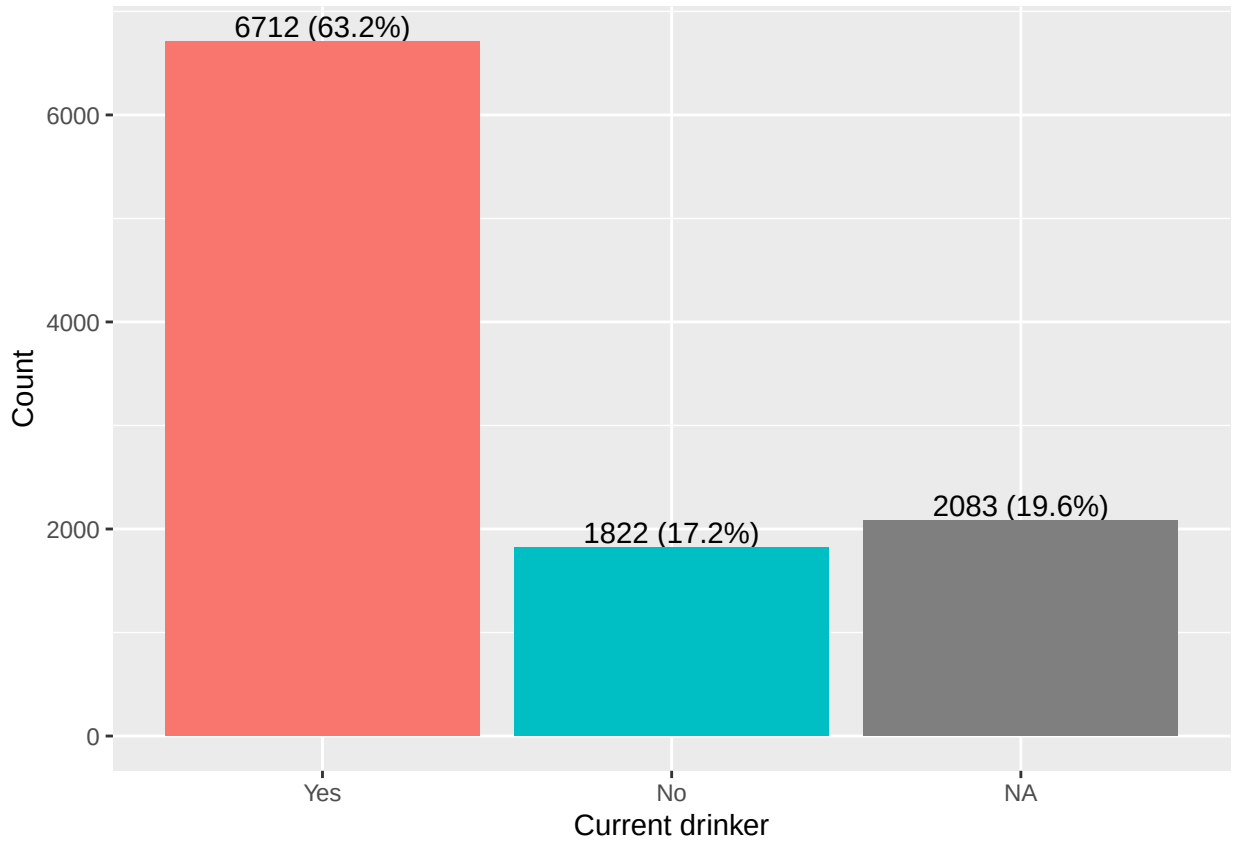



```
ggsave("Outputs/region_plot.png")
```

Current drinker

```
data%>%
  ggplot(aes(as_factor(drinks), fill=as_factor(drinks)))+
  geom_text(stat="count",aes(label=paste0(after_stat(count), " (", round(after_stat(count)/length(data$
            vjust=-0.2)+
```

```
  geom_bar()+
  labs(x="Current drinker",
       y="Count")+
  theme(legend.position = "none")->drinks_plot

drinks_plot
```




```
ggsave("Outputs/drinks_plot.png")
```

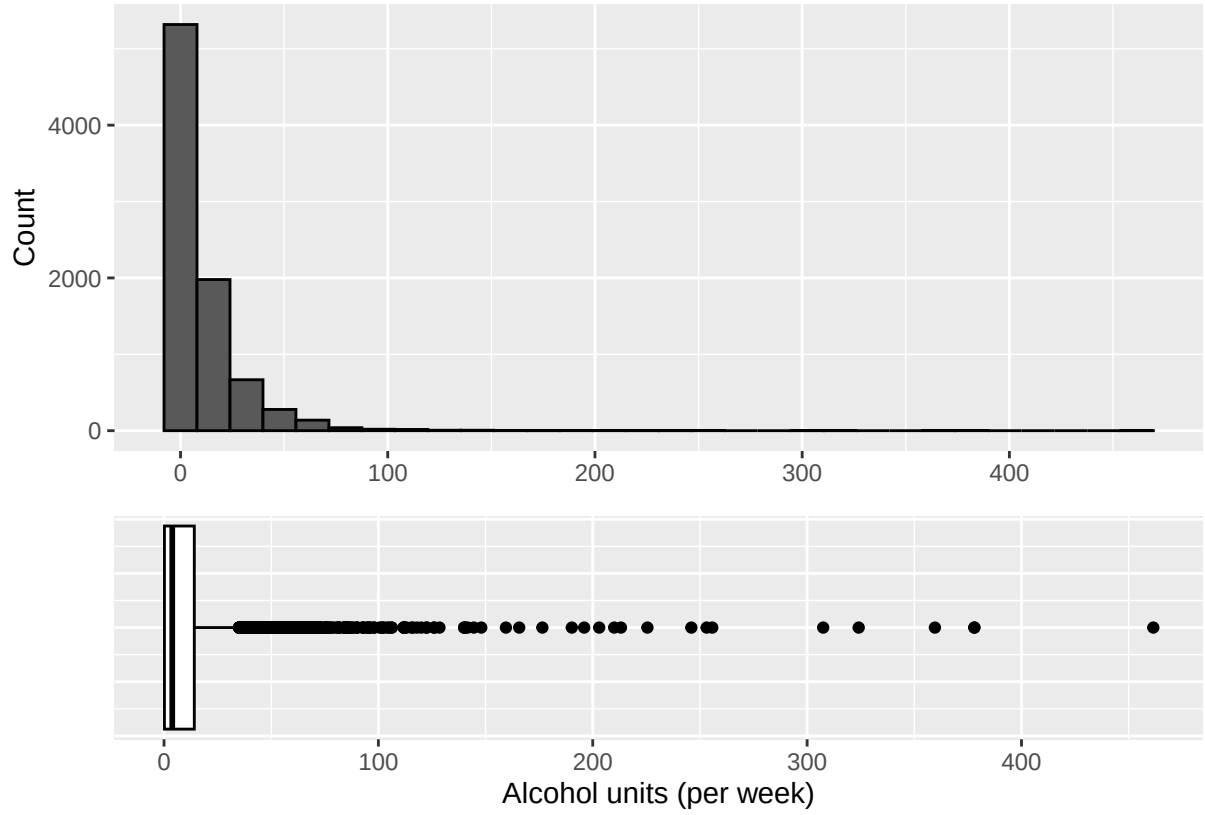
Units of alcohol per week

```
summary(data$alcohol_units)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##   0.0000   0.1875   3.8365  10.9088  14.1152 461.5000     2133
```

```
p1<-data%>%
  ggplot(aes(alcohol_units))+
  geom_histogram(color="black")+
  theme(axis.title.x = element_blank())+
  labs(y="Count")
```

```
p2<-data%>%
  ggplot(aes(alcohol_units))+
  geom_boxplot(color="black")+
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())+
  labs(x="Alcohol units (per week)")


p1/p2+plot_layout(nrow = 2, heights = c(2, 1))->alcohol_units_plot

alcohol_units_plot
```



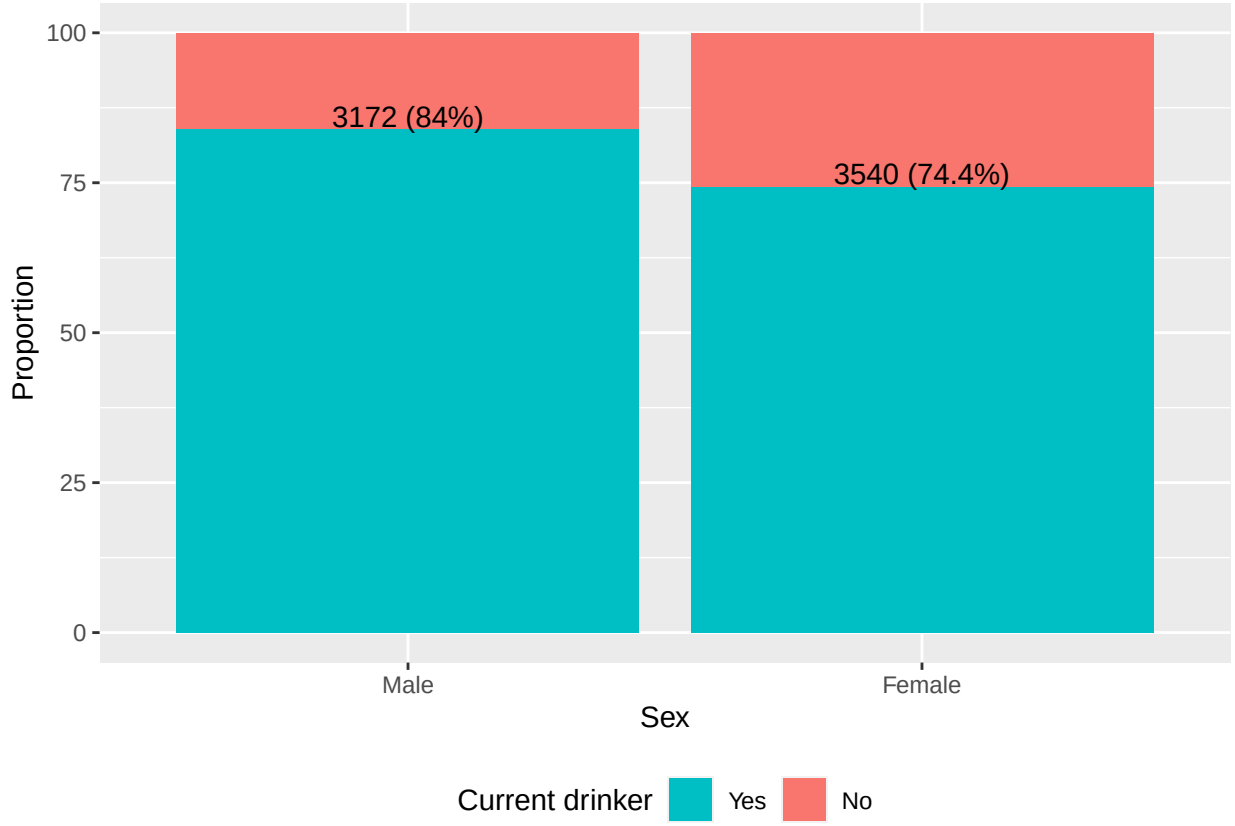

```
ggsave("Outputs/alcohol_units_plot.png")
```

Inferential statistics

Which gender drinks more alcohol

```
# plot (barchart)
data%>%
  filter(!is.na(drinks))%>%
  count(as_factor(sex), as_factor(drinks))%>%
```

```
  group_by(`as_factor(sex)`)%>%
  mutate(prop=round(n/sum(n)*100,1))%>%
  rename(Sex=`as_factor(sex)`,
         `Current drinker` = `as_factor(drinks)`)%>%

  ggplot(aes(Sex, y=prop,fill=fct_rev(`Current drinker`)))+
  geom_col()+
  labs(x="Sex",
       y="Proportion",
       fill="Current drinker")+
  guides(fill = guide_legend(reverse = TRUE))+
  geom_text(data=.%>%filter(`Current drinker`=="Yes"),
            aes(x=Sex, y=prop, label=paste0(n, " (", prop, "%)")),
            vjust=-0.1)+
  theme(legend.position = "bottom")
```




```
ggsave("Outputs/drinking_vs_sex.png",
       width=20,
       height=20,
       units="cm")

# contingency table

CrossTable(data$sex,
           data$drinks,
```

```
          prop.c = F,
prop.r = T,
prop.t=F,
prop.chisq = F,
chisq = F,
format="SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
##
## Total Observations in Table:  8534
##
##              | data$drinks
##     data$sex |         1  |         2  | Row Total |
## -------------|------------|------------|-----------|
##            1 |      3172  |       605  |      3777  |
##              |    83.982% |    16.018% |    44.258% |
## -------------|------------|------------|-----------|
##            2 |      3540  |      1217  |      4757  |
##              |    74.417% |    25.583% |    55.742% |
## -------------|------------|------------|-----------|
## Column Total |      6712  |      1822  |      8534  |
## -------------|------------|------------|-----------|
##
##
```

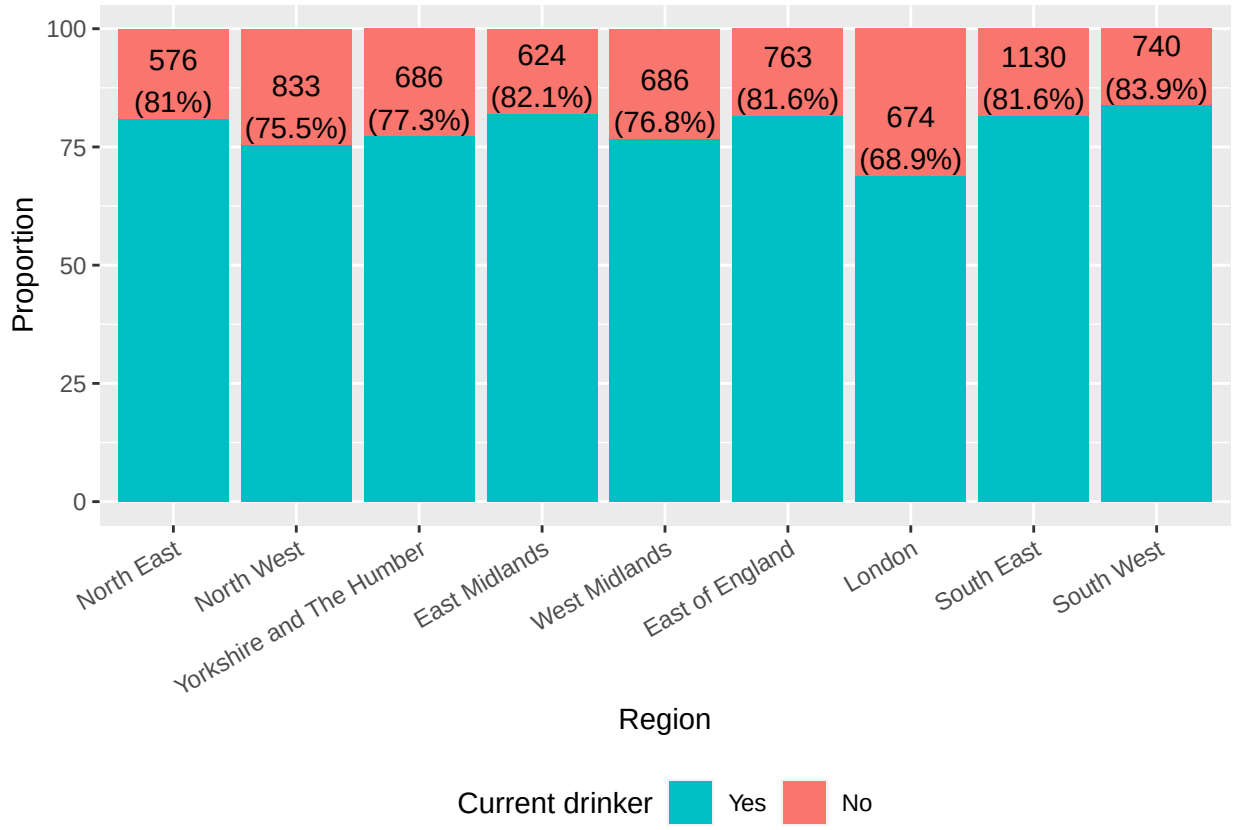
```
# run test

chisq.test(data$drinks, data$sex)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data$drinks and data$sex
## X-squared = 114.15, df = 1, p-value < 2.2e-16
```

Which region drinks more alcohol

```
# plot
data%>%
  filter(!is.na(drinks))%>%
  count(as_factor(region), as_factor(drinks))%>%
  group_by(`as_factor(region)`)%>%
  mutate(prop=round(n/sum(n)*100,1))%>%
  rename(Region=`as_factor(region)`,
         `Current drinker` = `as_factor(drinks)`)%>%
```

```
  ggplot(aes(Region, y=prop,fill=fct_rev(`Current drinker`)))+
  geom_col()+
  labs(x="Region",
       y="Proportion",
       fill="Current drinker")+
  guides(fill = guide_legend(reverse = TRUE))+
  geom_text(data=.%>%filter(`Current drinker`=="Yes"),
            aes(x=Region, y=prop, label=paste0(n, "\n(", prop, "%)")),
            vjust=-0.1)+
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle=30, hjust=1,vjust=1))
```




```
ggsave("Outputs/drinking_vs_region.png",
       width=20,
       height=20,
       units="cm")



# contingency table

CrossTable(data$region,
           data$drinks,
           prop.c = F,
prop.r = T,
```

```
prop.t=F,
prop.chisq = F,
chisq = F,
format="SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
##
## Total Observations in Table:  8534
##
##              | data$drinks
##  data$region |         1  |         2  | Row Total |
## -------------|------------|------------|-----------|
##            1 |       576  |       135  |       711  |
##              |    81.013% |    18.987% |     8.331% |
## -------------|------------|------------|-----------|
##            2 |       833  |       270  |      1103  |
##              |    75.521% |    24.479% |    12.925% |
## -------------|------------|------------|-----------|
##            3 |       686  |       201  |       887  |
##              |    77.339% |    22.661% |    10.394% |
## -------------|------------|------------|-----------|
##            4 |       624  |       136  |       760  |
##              |    82.105% |    17.895% |     8.906% |
## -------------|------------|------------|-----------|
##            5 |       686  |       207  |       893  |
##              |    76.820% |    23.180% |    10.464% |
## -------------|------------|------------|-----------|
##            6 |       763  |       172  |       935  |
##              |    81.604% |    18.396% |    10.956% |
## -------------|------------|------------|-----------|
##            7 |       674  |       304  |       978  |
##              |    68.916% |    31.084% |    11.460% |
## -------------|------------|------------|-----------|
##            8 |      1130  |       255  |      1385  |
##              |    81.588% |    18.412% |    16.229% |
## -------------|------------|------------|-----------|
##            9 |       740  |       142  |       882  |
##              |    83.900% |    16.100% |    10.335% |
## -------------|------------|------------|-----------|
## Column Total |      6712  |      1822  |      8534  |
## -------------|------------|------------|-----------|
##
##
```

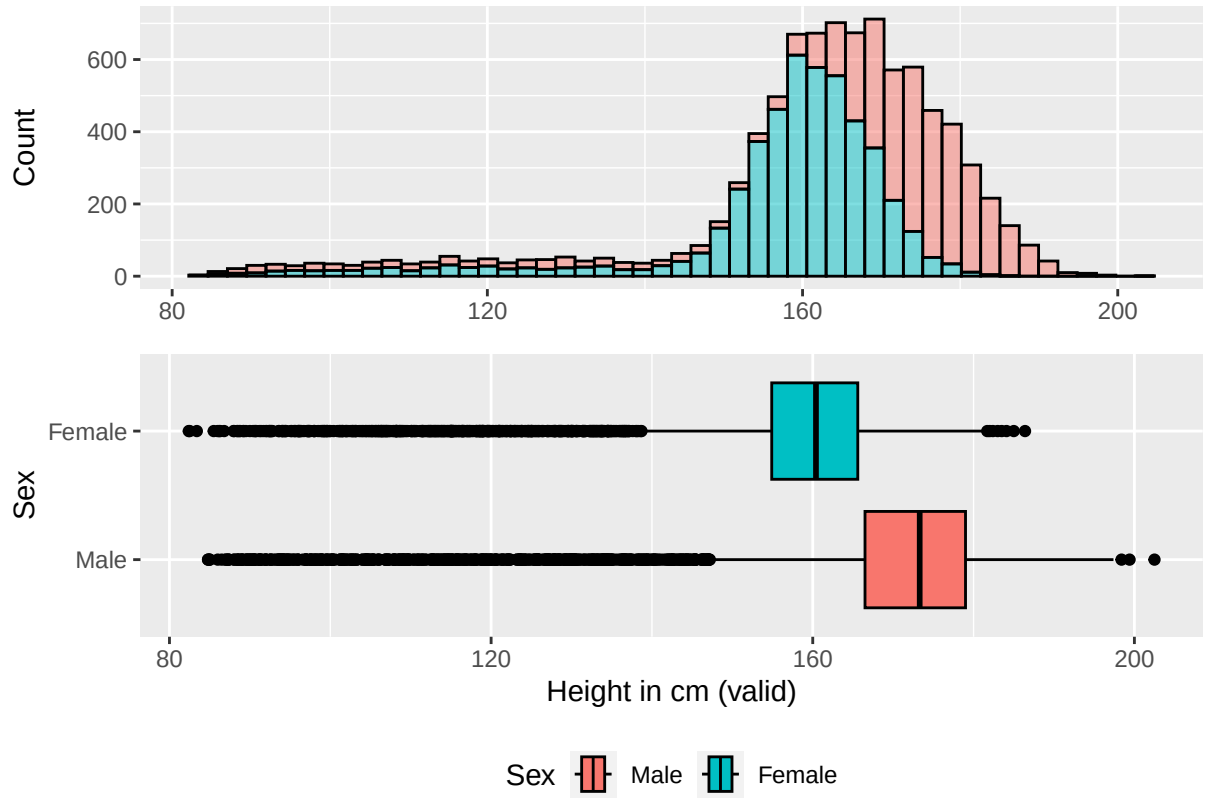
```
# run test
chisq.test(data$drinks, data$region)
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  data$drinks and data$region
## X-squared = 98.53, df = 8, p-value < 2.2e-16
```

Difference in height between men and women

```
# Plot
p1<-data%>%
  select(height, sex)%>%
  ggplot(aes(height, group=as_factor(sex), fill=as_factor(sex)))+
  geom_histogram(alpha=0.5,
                 color="black",
                 bins=50)+
  labs(x="Height in cm (valid)",
       y="Count",
       fill="Sex")+
  theme(legend.position = "none",
        axis.title.x = element_blank())

p2<-data%>%
  select(height, sex)%>%
  ggplot(aes(height, y=as_factor(sex),group=as_factor(sex), fill=as_factor(sex)))+
  geom_boxplot(color="black")+
  labs(x="Height in cm (valid)",
       y="Sex",
       fill="Sex")+
  theme(legend.position = "bottom")

p1/p2
```




```
ggsave("Outputs/height_by_sex.png",
       width=20,
       height=20,
       units="cm")


# test for normality

ks.test(data$height, "pnorm")
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  data$height
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
# Table
data%>%
  select(height, sex)%>%
  filter(!is.na(height))%>%
  group_by(as_factor(sex))%>%
  summarise(Minimum=min(height),
            Mean=round(mean(height),1),
            SD = round(sd(height),1),
```

```
            Median=round(median(height),1),
            Q1 = round(quantile(height, .25),1),
            Q3 = round(quantile(height, .75),1),
            Maximum=round(max(height),1))%>%
  rename(Sex=`as_factor(sex)`)->height_by_sex_summary_table


height_by_sex_summary_table
```

```
## # A tibble: 2 x 8
##   Sex    Minimum    Mean    SD Median    Q1    Q3 Maximum
##   <fct>  <dbl+lbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1 Male   84.8       167.  21     173.  166.  179     202.
## 2 Female 82.4       157.  15.4   160.  155.  166.    186.
```

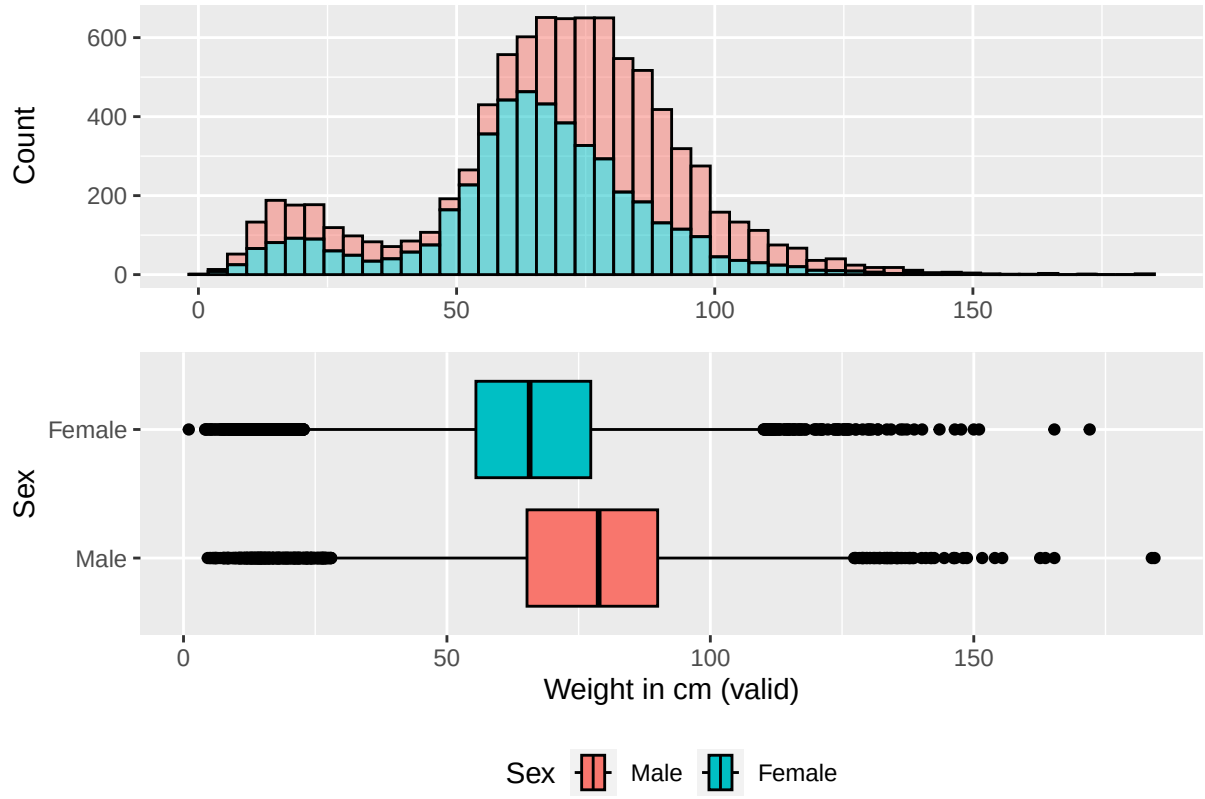
```
write_csv(height_by_sex_summary_table, "Outputs/height_by_sex_summary_table.csv")

# Test

## non-parametric wilcoxon independent samples test
wilcox.test(height~sex, data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  height by sex
## W = 14713021, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Difference in weight between men and women

```
# Plot
p1<-data%>%
  select(weight, sex)%>%
  ggplot(aes(weight, group=as_factor(sex), fill=as_factor(sex)))+
  geom_histogram(alpha=0.5,
                 color="black",
                 bins=50)+
  labs(x="Weight in kg (valid)",
       y="Count",
       fill="Sex")+
  theme(legend.position = "none",
        axis.title.x = element_blank())

p2<-data%>%
  select(weight, sex)%>%
  ggplot(aes(weight, y=as_factor(sex),group=as_factor(sex), fill=as_factor(sex)))+
  geom_boxplot(color="black")+
  labs(x="Weight in cm (valid)",
       y="Sex",
```

```
       fill="Sex")+
  theme(legend.position = "bottom")

p1/p2
```




```
ggsave("Outputs/weight_by_sex.png",
       width=20,
       height=20,
       units="cm")


# test for normality

ks.test(data$weight, "pnorm")
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  data$weight
## D = 0.99986, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
# Table
data%>%
```

```
  select(weight, sex)%>%
  filter(!is.na(weight))%>%
  group_by(as_factor(sex))%>%
  summarise(Minimum=min(weight),
            Mean=round(mean(weight),1),
            SD = round(sd(weight),1),
            Median=round(median(weight),1),
            Q1 = round(quantile(weight, .25),1),
            Q3 = round(quantile(weight, .75),1),
            Maximum=round(max(weight),1))%>%
  rename(Sex=`as_factor(sex)`)->weight_by_sex_summary_table


weight_by_sex_summary_table
```

```
## # A tibble: 2 x 8
##   Sex    Minimum    Mean    SD Median    Q1    Q3 Maximum
##   <fct>  <dbl+lbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1 Male   4.6        74.3    27   78.8  65.2  90      184.
## 2 Female 1          64.8    22   65.7  55.5  77.3    172
```

```
write_csv(weight_by_sex_summary_table, "Outputs/weight_by_sex_summary_table.csv")

# Test

## parametric independent samples  t-test
t.test(weight~sex, data, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  weight by sex
## t = 18.125, df = 8739, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##   8.479781 10.536397
## sample estimates:
## mean in group 1 mean in group 2
##        74.26612        64.75803
```

```
## non-parametric wilcoxon independent samples test
wilcox.test(weight~sex, data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  weight by sex
## W = 12449400, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

correlation between whether a person drinks nowadays, total household income, age at last birthday and gender?

```
mycor<- function(x,...){
  r<- apply(x, 2, function(j){
    apply(x, 2, function(i){
      as.numeric(cor.test(i,j,...)$estimate)
    })
  })
  P<- apply(x, 2, function(j){
    apply(x, 2, function(i){
      as.numeric(cor.test(i,j,...)$p.value)
    })
  })
  out<-c()
  out$P<- P
  out$r<- r
  return(out)
}

myCorDat<- mycor(data%>%
                   select(drinks, household_income, age, sex)%>%
                   mutate(drinks=case_when(drinks==1 ~ 1,
                                           drinks==2 ~ 0)),
                 method="pearson", na.action=na.omit)

myCorDat
```

```
## $P
##                        drinks household_income          age          sex
## drinks           0.000000e+00     2.644029e-11 1.871043e-10 6.231612e-27
## household_income 2.644029e-11     0.000000e+00 4.061712e-07 6.296692e-01
## age              1.871043e-10     4.061712e-07 0.000000e+00 7.558220e-04
## sex              6.231612e-27     6.296692e-01 7.558220e-04 0.000000e+00
##
## $r
##                       drinks household_income         age          sex
## drinks            1.00000000     -0.073253390 -0.06889968 -0.115941962
## household_income -0.07325339      1.000000000  0.04988733  0.004751272
## age              -0.06889968      0.049887330  1.00000000  0.032686539
## sex              -0.11594196      0.004751272  0.03268654  1.000000000
```