Datacamp_Cleaning Data in R_Preparing data for analysis

dizhen

2019/4/5

Type Conversion

```
as.character(2016)
```

```
## [1] "2016"
```

```
as.numeric(TRUE)
```

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

```
as.integer(99)
```

```
## [1] 99
```

```
as.factor("something")
```

```
## [1] something
## Levels: something
```

```
as.logical(0)
```

```
## [1] FALSE
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
# Experiment with basic lubridate functions
ymd("2015-08-25") # year-month-day
```

```
## [1] "2015-08-25"
```

```r
ymd("2015 August 25") # year-month-day
```

```
## [1] "2015-08-25"
```

```r
mdy("August 25, 2015") # month-day-year
```

```
## [1] "2015-08-25"
```

```r
hms("13:33:09") # hour-minute-second
```

```
## [1] "13H 33M 9S"
```

```r
ymd_hms("2015/08/25 13.33.09") # year-month-day hour-minute-second
```

```
## [1] "2015-08-25 13:33:09 UTC"
```

Practice

```r
# Make this evaluate to "character"
class("TRUE")
```

```
## [1] "character"
```

```r
# Make this evaluate to "numeric"
class(8484.00)
```

```
## [1] "numeric"
```

```r
# Make this evaluate to "integer"
class(99L)
```

```
## [1] "integer"
```

```r
# Make this evaluate to "factor"
class(factor("factor"))
```

```
## [1] "factor"
```

```r
# Make this evaluate to "logical"
class(FALSE)
```

```
## [1] "logical"
```

```r
library(readr)
students <- read_csv("data/students_with_dates.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   X1 = col_double(),
##   dob = col_date(format = ""),
##   Medu = col_double(),
##   Fedu = col_double(),
##   traveltime = col_double(),
##   studytime = col_double(),
##   failures = col_double(),
##   famrel = col_double(),
##   freetime = col_double(),
##   goout = col_double(),
##   Dalc = col_double(),
##   Walc = col_double(),
##   health = col_double(),
##   nurse_visit = col_datetime(format = ""),
##   absences = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
# Preview students with str()
str(students)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 395 obs. of  33 variables:
##  $ X1         : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ school     : chr  "GP" "GP" "GP" "GP" ...
##  $ sex        : chr  "F" "F" "F" "F" ...
##  $ dob        : Date, format: "2000-06-05" "1999-11-25" ...
##  $ address    : chr  "U" "U" "U" "U" ...
##  $ famsize    : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus    : chr  "A" "T" "T" "T" ...
##  $ Medu       : num  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu       : num  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob       : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob       : chr  "teacher" "other" "other" "services" ...
##  $ reason     : chr  "course" "course" "other" "home" ...
##  $ guardian   : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime : num  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime  : num  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures   : num  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup  : chr  "yes" "no" "yes" "no" ...
##  $ famsup     : chr  "no" "yes" "no" "yes" ...
##  $ paid       : chr  "no" "no" "yes" "yes" ...
##  $ activities : chr  "no" "no" "no" "yes" ...
##  $ nursery    : chr  "yes" "no" "yes" "yes" ...
##  $ higher     : chr  "yes" "yes" "yes" "yes" ...
##  $ internet   : chr  "no" "yes" "yes" "yes" ...
##  $ romantic   : chr  "no" "no" "no" "yes" ...
##  $ famrel     : num  4 5 4 3 4 5 4 4 4 5 ...
```

```
##  $ freetime   : num  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout      : num  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc       : num  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc       : num  1 1 3 1 2 2 1 1 1 1 ...
##  $ health     : num  3 3 3 5 5 5 3 1 1 5 ...
##  $ nurse_visit: POSIXct, format: "2014-04-10 14:59:54" "2015-03-12 14:59:54" ...
##  $ absences   : num  6 4 10 2 4 10 0 6 0 0 ...
##  $ Grades     : chr  "5/6/6" "5/5/6" "7/8/10" "15/14/15" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   X1 = col_double(),
##   ..   school = col_character(),
##   ..   sex = col_character(),
##   ..   dob = col_date(format = ""),
##   ..   address = col_character(),
##   ..   famsize = col_character(),
##   ..   Pstatus = col_character(),
##   ..   Medu = col_double(),
##   ..   Fedu = col_double(),
##   ..   Mjob = col_character(),
##   ..   Fjob = col_character(),
##   ..   reason = col_character(),
##   ..   guardian = col_character(),
##   ..   traveltime = col_double(),
##   ..   studytime = col_double(),
##   ..   failures = col_double(),
##   ..   schoolsup = col_character(),
##   ..   famsup = col_character(),
##   ..   paid = col_character(),
##   ..   activities = col_character(),
##   ..   nursery = col_character(),
##   ..   higher = col_character(),
##   ..   internet = col_character(),
##   ..   romantic = col_character(),
##   ..   famrel = col_double(),
##   ..   freetime = col_double(),
##   ..   goout = col_double(),
##   ..   Dalc = col_double(),
##   ..   Walc = col_double(),
##   ..   health = col_double(),
##   ..   nurse_visit = col_datetime(format = ""),
##   ..   absences = col_double(),
##   ..   Grades = col_character()
##   .. )
```

```r
# Coerce Grades to character
students$Grades <- as.character(students$Grades)

# Coerce Medu to factor
students$Medu <- as.factor(students$Medu)

# Coerce Fedu to factor
students$Fedu <- as.factor(students$Fedu)
```

```r
# Look at students once more with str()
str(students)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 395 obs. of  33 variables:
##  $ X1         : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ school     : chr  "GP" "GP" "GP" "GP" ...
##  $ sex        : chr  "F" "F" "F" "F" ...
##  $ dob        : Date, format: "2000-06-05" "1999-11-25" ...
##  $ address    : chr  "U" "U" "U" "U" ...
##  $ famsize    : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus    : chr  "A" "T" "T" "T" ...
##  $ Medu       : Factor w/ 5 levels "0","1","2","3",..: 5 2 2 5 4 5 3 5 4 4 ...
##  $ Fedu       : Factor w/ 5 levels "0","1","2","3",..: 5 2 2 3 4 4 3 5 3 5 ...
##  $ Mjob       : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob       : chr  "teacher" "other" "other" "services" ...
##  $ reason     : chr  "course" "course" "other" "home" ...
##  $ guardian   : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime : num  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime  : num  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures   : num  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup  : chr  "yes" "no" "yes" "no" ...
##  $ famsup     : chr  "no" "yes" "no" "yes" ...
##  $ paid       : chr  "no" "no" "yes" "yes" ...
##  $ activities : chr  "no" "no" "no" "yes" ...
##  $ nursery    : chr  "yes" "no" "yes" "yes" ...
##  $ higher     : chr  "yes" "yes" "yes" "yes" ...
##  $ internet   : chr  "no" "yes" "yes" "yes" ...
##  $ romantic   : chr  "no" "no" "no" "yes" ...
##  $ famrel     : num  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime   : num  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout      : num  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc       : num  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc       : num  1 1 3 1 2 2 1 1 1 1 ...
##  $ health     : num  3 3 3 5 5 5 3 1 1 5 ...
##  $ nurse_visit: POSIXct, format: "2014-04-10 14:59:54" "2015-03-12 14:59:54" ...
##  $ absences   : num  6 4 10 2 4 10 0 6 0 0 ...
##  $ Grades     : chr  "5/6/6" "5/5/6" "7/8/10" "15/14/15" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   X1 = col_double(),
##   ..   school = col_character(),
##   ..   sex = col_character(),
##   ..   dob = col_date(format = ""),
##   ..   address = col_character(),
##   ..   famsize = col_character(),
##   ..   Pstatus = col_character(),
##   ..   Medu = col_double(),
##   ..   Fedu = col_double(),
##   ..   Mjob = col_character(),
##   ..   Fjob = col_character(),
##   ..   reason = col_character(),
##   ..   guardian = col_character(),
##   ..   traveltime = col_double(),
```

```
##   ..   studytime = col_double(),
##   ..   failures = col_double(),
##   ..   schoolsup = col_character(),
##   ..   famsup = col_character(),
##   ..   paid = col_character(),
##   ..   activities = col_character(),
##   ..   nursery = col_character(),
##   ..   higher = col_character(),
##   ..   internet = col_character(),
##   ..   romantic = col_character(),
##   ..   famrel = col_double(),
##   ..   freetime = col_double(),
##   ..   goout = col_double(),
##   ..   Dalc = col_double(),
##   ..   Walc = col_double(),
##   ..   health = col_double(),
##   ..   nurse_visit = col_datetime(format = ""),
##   ..   absences = col_double(),
##   ..   Grades = col_character()
##   .. )
```

```
# Preview students2 with str()
students2 <- students

# Load the lubridate package
library(lubridate)

# Parse as date
dmy("17 Sep 2015")
```

```
## [1] "2015-09-17"
```

```
# Parse as date and time (with no seconds!)
mdy_hm("July 15, 2012 12:56")
```

```
## [1] "2012-07-15 12:56:00 UTC"
```

```
# Coerce dob to a date (with no time)
students2$dob <- ymd(students2$dob)

# Coerce nurse_visit to a date and time
students2$nurse_visit <- ymd_hms(students2$nurse_visit)

# Look at students2 once more with str()
str(students2)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 395 obs. of  33 variables:
##  $ X1         : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ school     : chr  "GP" "GP" "GP" "GP" ...
##  $ sex        : chr  "F" "F" "F" "F" ...
##  $ dob        : Date, format: "2000-06-05" "1999-11-25" ...
##  $ address    : chr  "U" "U" "U" "U" ...
```

```
##  $ famsize    : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus    : chr  "A" "T" "T" "T" ...
##  $ Medu       : Factor w/ 5 levels "0","1","2","3",..: 5 2 2 5 4 5 3 5 4 4 ...
##  $ Fedu       : Factor w/ 5 levels "0","1","2","3",..: 5 2 2 3 4 4 3 5 3 5 ...
##  $ Mjob       : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob       : chr  "teacher" "other" "other" "services" ...
##  $ reason     : chr  "course" "course" "other" "home" ...
##  $ guardian   : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime : num  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime  : num  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures   : num  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup  : chr  "yes" "no" "yes" "no" ...
##  $ famsup     : chr  "no" "yes" "no" "yes" ...
##  $ paid       : chr  "no" "no" "yes" "yes" ...
##  $ activities : chr  "no" "no" "no" "yes" ...
##  $ nursery    : chr  "yes" "no" "yes" "yes" ...
##  $ higher     : chr  "yes" "yes" "yes" "yes" ...
##  $ internet   : chr  "no" "yes" "yes" "yes" ...
##  $ romantic   : chr  "no" "no" "no" "yes" ...
##  $ famrel     : num  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime   : num  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout      : num  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc       : num  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc       : num  1 1 3 1 2 2 1 1 1 1 ...
##  $ health     : num  3 3 3 5 5 5 3 1 1 5 ...
##  $ nurse_visit: POSIXct, format: "2014-04-10 14:59:54" "2015-03-12 14:59:54" ...
##  $ absences   : num  6 4 10 2 4 10 0 6 0 0 ...
##  $ Grades     : chr  "5/6/6" "5/5/6" "7/8/10" "15/14/15" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   X1 = col_double(),
##   ..   school = col_character(),
##   ..   sex = col_character(),
##   ..   dob = col_date(format = ""),
##   ..   address = col_character(),
##   ..   famsize = col_character(),
##   ..   Pstatus = col_character(),
##   ..   Medu = col_double(),
##   ..   Fedu = col_double(),
##   ..   Mjob = col_character(),
##   ..   Fjob = col_character(),
##   ..   reason = col_character(),
##   ..   guardian = col_character(),
##   ..   traveltime = col_double(),
##   ..   studytime = col_double(),
##   ..   failures = col_double(),
##   ..   schoolsup = col_character(),
##   ..   famsup = col_character(),
##   ..   paid = col_character(),
##   ..   activities = col_character(),
##   ..   nursery = col_character(),
##   ..   higher = col_character(),
##   ..   internet = col_character(),
##   ..   romantic = col_character(),
```

```
##   ..   famrel = col_double(),
##   ..   freetime = col_double(),
##   ..   goout = col_double(),
##   ..   Dalc = col_double(),
##   ..   Walc = col_double(),
##   ..   health = col_double(),
##   ..   nurse_visit = col_datetime(format = ""),
##   ..   absences = col_double(),
##   ..   Grades = col_character()
##   .. )
```

String manipulation

1. Key functions in stringr for cleaning data

str_trim() - Trim leading and trailing white space

str_pad() - Pad with additional characters

str_detect() - Detect a pa!ern

str_replace() - Find and replace a pa!ern

```
library("stringr")
# Trim leading and trailing white space
str_trim(" this is a test ")
```

```
## [1] "this is a test"
```

```
# Pad string with zeros
str_pad("24493", width = 7, side = "left", pad = "0")
```

```
## [1] "0024493"
```

```
# Create character vector of names
friends <- c("Sarah", "Tom", "Alice")

# Search for string in vector
str_detect(friends, "Alice")
```

```
## [1] FALSE FALSE  TRUE
```

```
# Replace string in vector
str_replace(friends, "Alice", "David")
```

```
## [1] "Sarah" "Tom"   "David"
```

2. Other helpful functions in base R

tolower() - Make all lowercase

toupper() - Make all uppercase

```r
# Make all lowercase
tolower("I AM TALKING LOUDLY!!")
```

```
## [1] "i am talking loudly!!"
```

```r
# Make all uppercase
toupper("I am whispering...")
```

```
## [1] "I AM WHISPERING..."
```

```r
# Load the stringr package
library(stringr)

# Trim all leading and trailing whitespace
str_trim(c("   Filip ", "Nick  ", " Jonathan"))
```

```
## [1] "Filip"    "Nick"     "Jonathan"
```

```r
# Pad these strings with leading zeros
str_pad(c("23485W", "8823453Q", "994Z"),width = 9, side = 'left', pad = '0')
```

```
## [1] "00023485W" "08823453Q" "00000994Z"
```

```r
# Print state abbreviations
states<- c("al", "ak", "az", "ar", "ca", "co", "ct", "de", "fl", "ga", "hi", "id")

# Make states all uppercase and save result to states_upper
states_upper <- toupper(states)
states_upper
```

```
##  [1] "AL" "AK" "AZ" "AR" "CA" "CO" "CT" "DE" "FL" "GA" "HI" "ID"
```

```r
# Make states_upper all lowercase again
tolower(states_upper)
```

```
##  [1] "al" "ak" "az" "ar" "ca" "co" "ct" "de" "fl" "ga" "hi" "id"
```

```r
# Copy of students2: students3
students3 <- students2

# Look at the head of students3
head(students3)
```

```
## # A tibble: 6 x 33
##      X1 school sex   dob        address famsize Pstatus Medu  Fedu  Mjob
##   <dbl> <chr>  <chr> <date>     <chr>   <chr>   <chr>   <fct> <fct> <chr>
## 1     1 GP     F     2000-06-05 U       GT3     A       4     4     at_h~
## 2     2 GP     F     1999-11-25 U       GT3     T       1     1     at_h~
## 3     3 GP     F     1998-02-02 U       LE3     T       1     1     at_h~
```

```
## 4     4 GP     F     1997-12-20 U       GT3     T       4     2     heal~
## 5     5 GP     F     1998-10-04 U       GT3     T       3     3     other
## 6     6 GP     M     1999-06-16 U       LE3     T       4     3     serv~
## # ... with 23 more variables: Fjob <chr>, reason <chr>, guardian <chr>,
## #   traveltime <dbl>, studytime <dbl>, failures <dbl>, schoolsup <chr>,
## #   famsup <chr>, paid <chr>, activities <chr>, nursery <chr>,
## #   higher <chr>, internet <chr>, romantic <chr>, famrel <dbl>,
## #   freetime <dbl>, goout <dbl>, Dalc <dbl>, Walc <dbl>, health <dbl>,
## #   nurse_visit <dttm>, absences <dbl>, Grades <chr>
```

```
# Detect all dates of birth (dob) in 1997
str_detect(students3$dob,'1997')
```

```
##   [1] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE
##  [12] FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
##  [23]  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
##  [34] FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE
##  [45] FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE
##  [56] FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [67] FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
##  [78] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE
##  [89] FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## [100]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [111] FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
## [122]  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
## [133] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE
## [144]  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE
## [155] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE
## [166]  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE
## [177] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
## [188] FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [199] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [210]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE
## [221] FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
## [232] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## [243] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## [254] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [265] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [276] FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
## [287] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## [298]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [309] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [320] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE
## [331] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
## [342] FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE
## [353] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE
## [364] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## [375]  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
## [386] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
# In the sex column, replace "F" with "Female" ...
students3$sex <- str_replace(students3$sex,'F','Female')
```

```r
# ... and "M" with "Male"
students3$sex <- str_replace(students3$sex,'M','Male')

# View the head of students3
head(students3)
```

```
## # A tibble: 6 x 33
##      X1 school sex   dob        address famsize Pstatus Medu  Fedu  Mjob
##   <dbl> <chr>  <chr> <date>     <chr>   <chr>   <chr>   <fct> <fct> <chr>
## 1     1 GP     Fema~ 2000-06-05 U       GT3     A       4     4     at_h~
## 2     2 GP     Fema~ 1999-11-25 U       GT3     T       1     1     at_h~
## 3     3 GP     Fema~ 1998-02-02 U       LE3     T       1     1     at_h~
## 4     4 GP     Fema~ 1997-12-20 U       GT3     T       4     2     heal~
## 5     5 GP     Fema~ 1998-10-04 U       GT3     T       3     3     other
## 6     6 GP     Male  1999-06-16 U       LE3     T       4     3     serv~
## # ... with 23 more variables: Fjob <chr>, reason <chr>, guardian <chr>,
## #   traveltime <dbl>, studytime <dbl>, failures <dbl>, schoolsup <chr>,
## #   famsup <chr>, paid <chr>, activities <chr>, nursery <chr>,
## #   higher <chr>, internet <chr>, romantic <chr>, famrel <dbl>,
## #   freetime <dbl>, goout <dbl>, Dalc <dbl>, Walc <dbl>, health <dbl>,
## #   nurse_visit <dttm>, absences <dbl>, Grades <chr>
```

Missing and special values

1. Missing values

May be random, but dangerous to assume

Sometimes associated with variable/outcome of interest

In R, represented as NA

May appear in other forms

#N/A (Excel), Single dot (SPSS, SAS), Empty string

2. Special values

Inf - “Infinite value” (indicative of outliers?)

NaN - “Not a number” (rethink a variable?)

3. Finding missing values

```r
# Create small dataset
df <- data.frame(A = c(1, NA, 8, NA), B = c(3, NA, 88, 23), C = c(2, 45, 3, 1))

# Check for NAs
is.na(df)
```

```
##          A     B     C
## [1,] FALSE FALSE FALSE
## [2,]  TRUE  TRUE FALSE
## [3,] FALSE FALSE FALSE
## [4,]  TRUE FALSE FALSE
```

```r
# Are there any NAs?
any(is.na(df))
```

```
## [1] TRUE
```

```r
# Count number of NAs
sum(is.na(df))
```

```
## [1] 3
```

```r
# Use summary() to find NAs
summary(df)
```

```
##        A               B               C
##  Min.   :1.00   Min.   : 3.0   Min.   : 1.00
##  1st Qu.:2.75   1st Qu.:13.0   1st Qu.: 1.75
##  Median :4.50   Median :23.0   Median : 2.50
##  Mean   :4.50   Mean   :38.0   Mean   :12.75
##  3rd Qu.:6.25   3rd Qu.:55.5   3rd Qu.:13.50
##  Max.   :8.00   Max.   :88.0   Max.   :45.00
##  NA's   :2      NA's   :1
```

4. Dealing with missing values

```r
# Find rows with no missing values
complete.cases(df)
```

```
## [1]  TRUE FALSE  TRUE FALSE
```

```r
# Subset data, keeping only complete cases
df[complete.cases(df), ]
```

```
##   A  B C
## 1 1  3 2
## 3 8 88 3
```

```r
# Another way to remove rows with NAs
na.omit(df)
```

```
##   A  B C
## 1 1  3 2
## 3 8 88 3
```

Practice

```r
social_df <- data.frame(name = factor(c("Sarah", "Tom","David", "Alice")), n_friends = c(244, NA, 145, 4

# Call is.na() on the full social_df to spot all NAs
is.na(social_df)
```

```
##       name n_friends status
## [1,] FALSE     FALSE  FALSE
## [2,] FALSE      TRUE  FALSE
## [3,] FALSE     FALSE  FALSE
## [4,] FALSE     FALSE  FALSE
```

```
# Use the any() function to ask whether there are any NAs in the data
any(is.na(social_df))
```

```
## [1] TRUE
```

```
# View a summary() of the dataset
summary(social_df)
```

```
##     name     n_friends                 status
##  Alice:1   Min.   : 43.0                  :2
##  David:1   1st Qu.: 94.0   Going out!     :1
##  Sarah:1   Median :145.0   Moving night...:1
##  Tom  :1   Mean   :144.0
##            3rd Qu.:194.5
##            Max.   :244.0
##            NA's   :1
```

```
# Call table() on the status column
table(social_df$status)
```

```
##
##                      Going out! Moving night...
##               2               1               1
```

```
# Replace all empty strings in status with NA
social_df$status[social_df$status == ""] <- NA

# Print social_df to the console
social_df
```

```
##    name n_friends          status
## 1 Sarah       244      Going out!
## 2   Tom        NA            <NA>
## 3 David       145 Moving night...
## 4 Alice        43            <NA>
```

```
# Use complete.cases() to see which rows have no missing values
complete.cases(social_df)
```

```
## [1]  TRUE FALSE  TRUE FALSE
```

```
# Use na.omit() to remove all rows with any missing values
na.omit(social_df)
```

```
##    name n_friends          status
## 1 Sarah       244      Going out!
## 3 David       145 Moving night...
```

Outliers and obvious errors

1. outliers

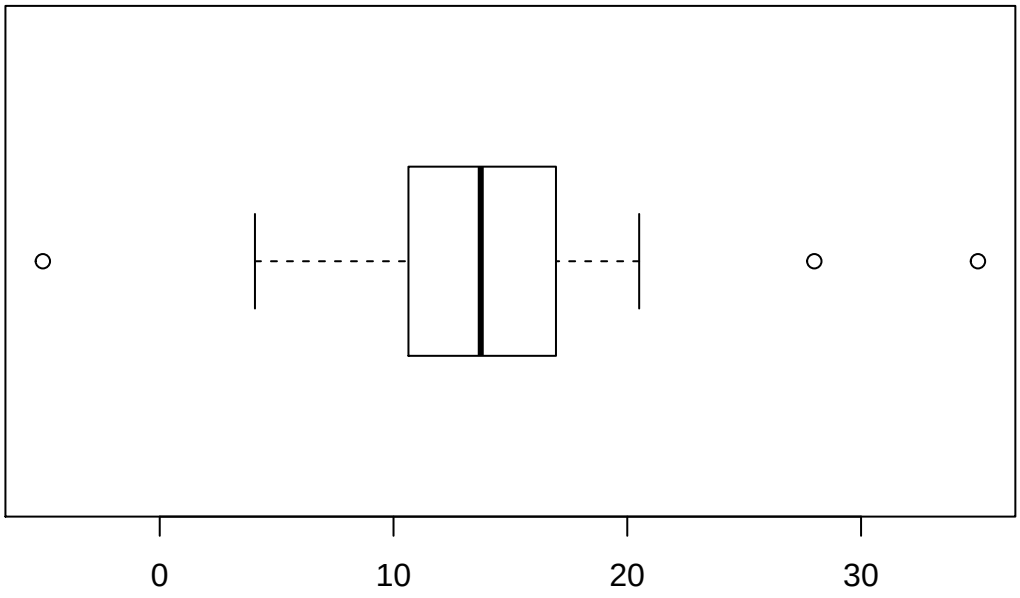
Extreme values distant from other values

Several causes

-Valid measurements, Variability in measurement, Experimental error, Data entry error

May be discarded or retained depending on cause

```
# Simulate some data
set.seed(10)
x <- c(rnorm(30, mean = 15, sd = 5), -5, 28, 35)

# View a boxplot
boxplot(x, horizontal = TRUE)
```

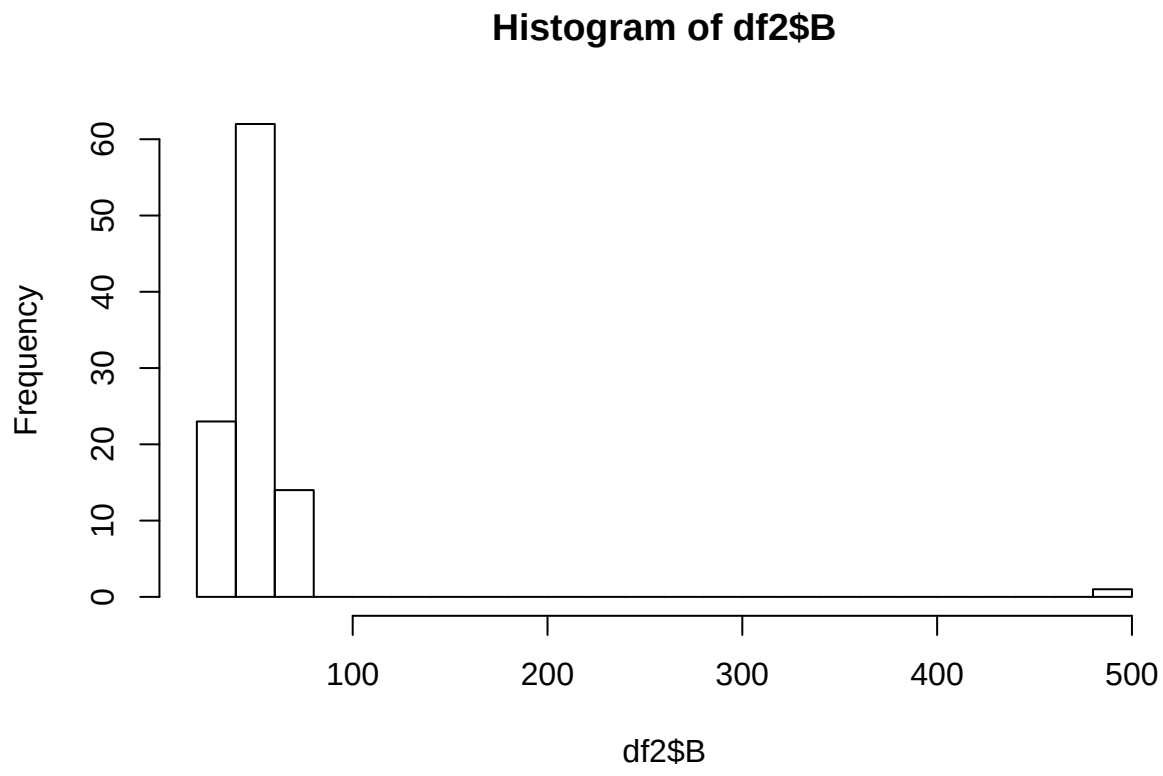


2. Finding outliers and errors

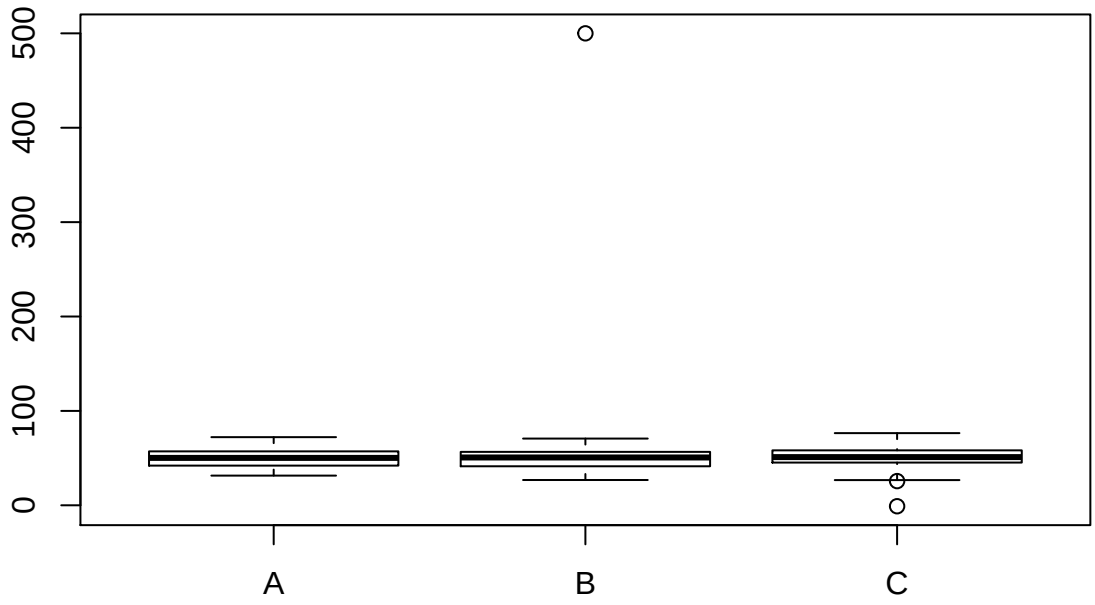
```
# Create another small dataset
df2 <- data.frame(A = rnorm(100, 50, 10), B = c(rnorm(99, 50, 10), 500), C = c(rnorm(99, 50, 10), -1))
# View a summary
summary(df2)
```

```
##        A                B                C         
##  Min.   :31.46   Min.   : 26.79   Min.   :-1.00  
##  1st Qu.:42.21   1st Qu.: 41.35   1st Qu.:45.29  
##  Median :50.20   Median : 50.67   Median :51.06  
##  Mean   :49.70   Mean   : 53.62   Mean   :50.88  
##  3rd Qu.:57.12   3rd Qu.: 56.57   3rd Qu.:58.13  
##  Max.   :72.21   Max.   :500.00   Max.   :76.44
```

```r
# View a histogram
hist(df2$B, breaks = 20)
```




```r
# View a boxplot
boxplot(df2)
```



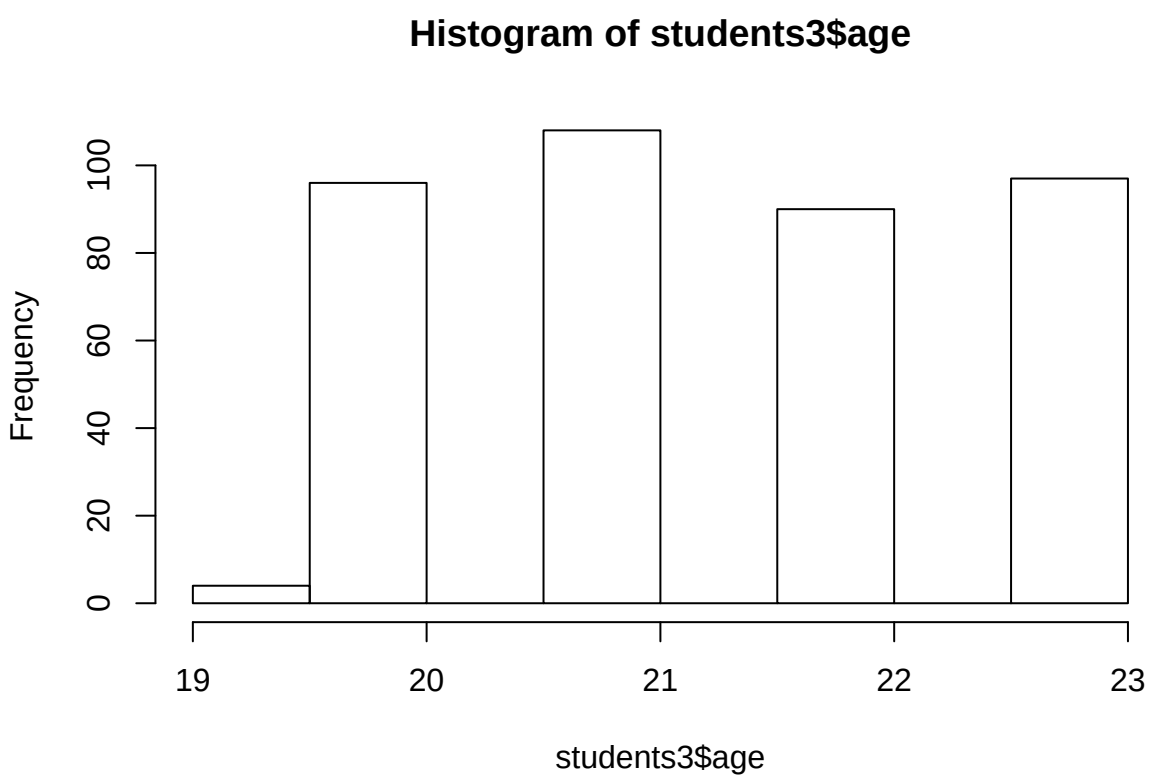
Practice

```
# Look at a summary() of students3
summary(students3)
```

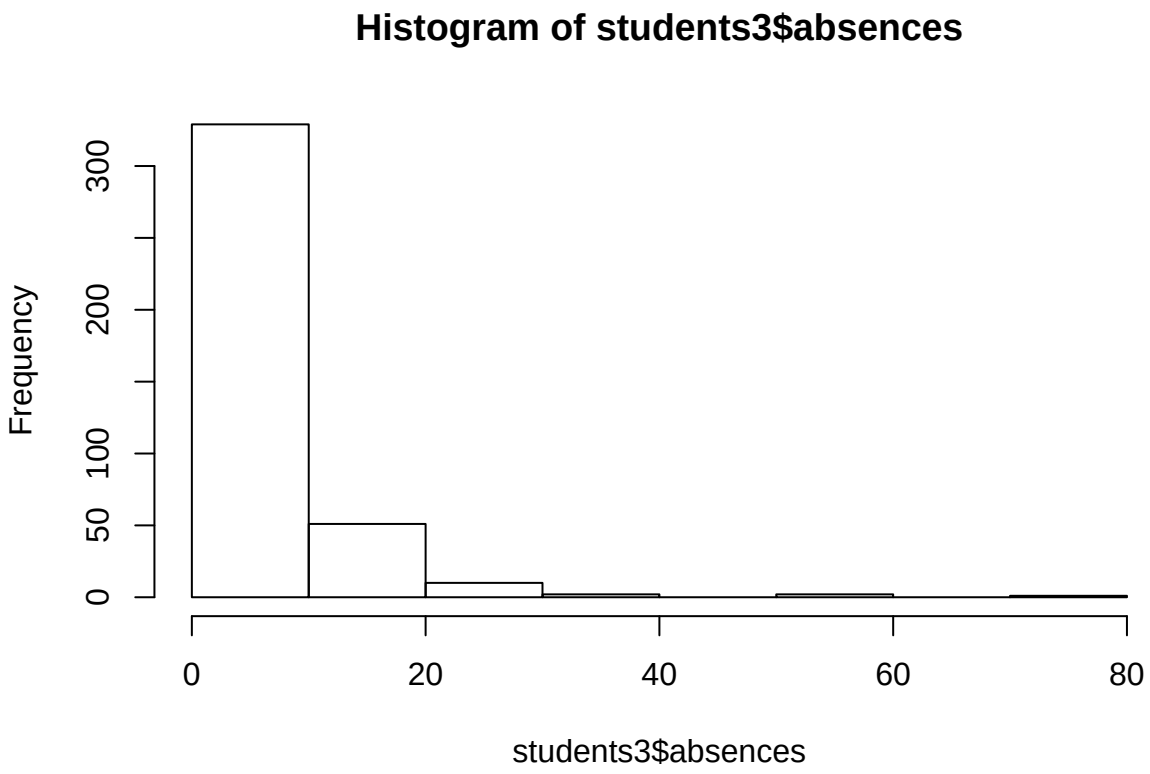
```
##        X1           school              sex
##  Min.   :  1.0   Length:395         Length:395
##  1st Qu.: 99.5   Class :character   Class :character
##  Median :198.0   Mode  :character   Mode  :character
##  Mean   :198.0
##  3rd Qu.:296.5
##  Max.   :395.0
##       dob               address            famsize
##  Min.   :1996-11-02   Length:395         Length:395
##  1st Qu.:1997-11-04   Class :character   Class :character
##  Median :1998-12-16   Mode  :character   Mode  :character
##  Mean   :1998-10-30
##  3rd Qu.:1999-10-29
##  Max.   :2000-10-25
##    Pstatus          Medu    Fedu        Mjob               Fjob
##  Length:395         0:  3   0:  2   Length:395         Length:395
##  Class :character   1: 59   1: 82   Class :character   Class :character
##  Mode  :character   2:103   2:115   Mode  :character   Mode  :character
##                     3: 99   3:100
```

```
##                     4:131   4: 96
##
##     reason            guardian           traveltime      studytime
##  Length:395         Length:395         Min.   :1.000   Min.   :1.000
##  Class :character   Class :character   1st Qu.:1.000   1st Qu.:1.000
##  Mode  :character   Mode  :character   Median :1.000   Median :2.000
##                                        Mean   :1.448   Mean   :2.035
##                                        3rd Qu.:2.000   3rd Qu.:2.000
##                                        Max.   :4.000   Max.   :4.000
##     failures       schoolsup           famsup              paid
##  Min.   :0.0000   Length:395         Length:395         Length:395
##  1st Qu.:0.0000   Class :character   Class :character   Class :character
##  Median :0.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.3342
##  3rd Qu.:0.0000
##  Max.   :3.0000
##   activities          nursery            higher
##  Length:395         Length:395         Length:395
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##    internet           romantic            famrel         freetime
##  Length:395         Length:395         Min.   :1.000   Min.   :1.000
##  Class :character   Class :character   1st Qu.:4.000   1st Qu.:3.000
##  Mode  :character   Mode  :character   Median :4.000   Median :3.000
##                                        Mean   :3.944   Mean   :3.235
##                                        3rd Qu.:5.000   3rd Qu.:4.000
##                                        Max.   :5.000   Max.   :5.000
##      goout            Dalc            Walc           health
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000
##  Median :3.000   Median :1.000   Median :2.000   Median :4.000
##  Mean   :3.109   Mean   :1.481   Mean   :2.291   Mean   :3.554
##  3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##   nurse_visit                     absences         Grades
##  Min.   :2013-10-28 14:59:54   Min.   : 0.000   Length:395
##  1st Qu.:2014-04-07 02:59:54   1st Qu.: 0.000   Class :character
##  Median :2014-09-15 14:59:54   Median : 4.000   Mode  :character
##  Mean   :2014-10-10 05:31:11   Mean   : 5.709
##  3rd Qu.:2015-04-08 02:59:54   3rd Qu.: 8.000
##  Max.   :2015-10-15 14:59:54   Max.   :75.000
```

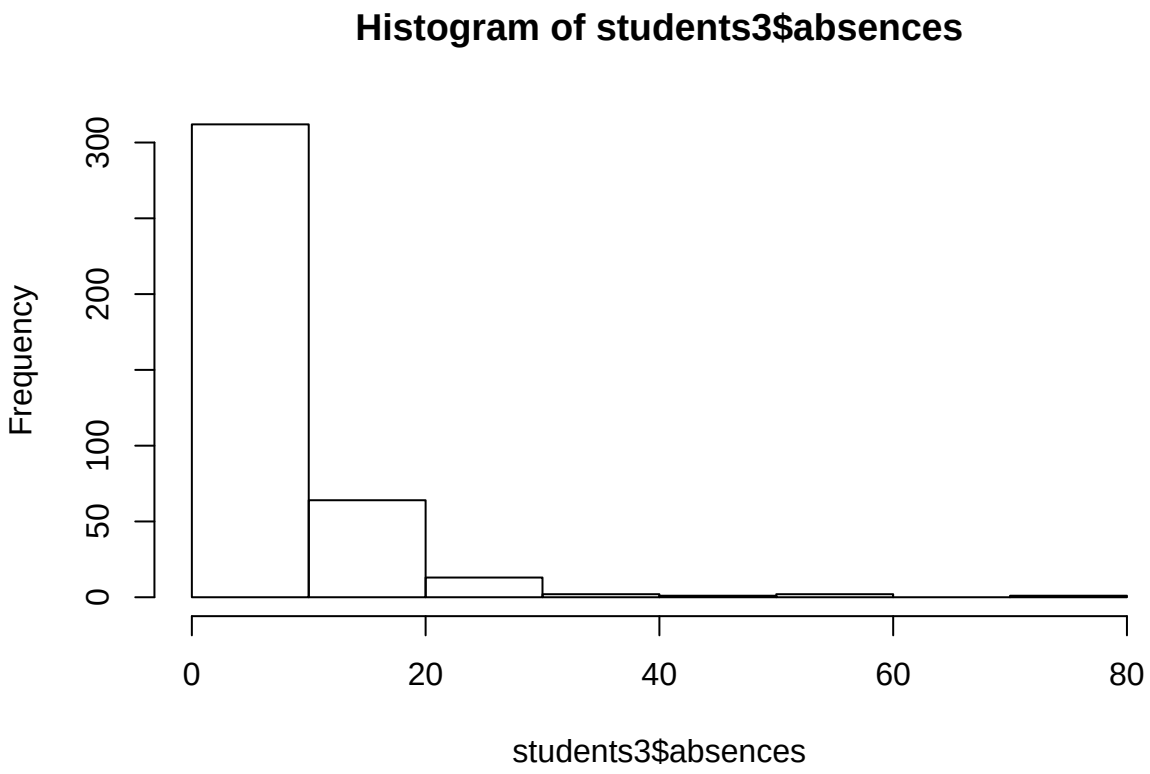
```r
students3$age <- as.numeric(round((Sys.Date() - students3$dob)/365))

# View a histogram of the age variable
hist(students3$age)
```



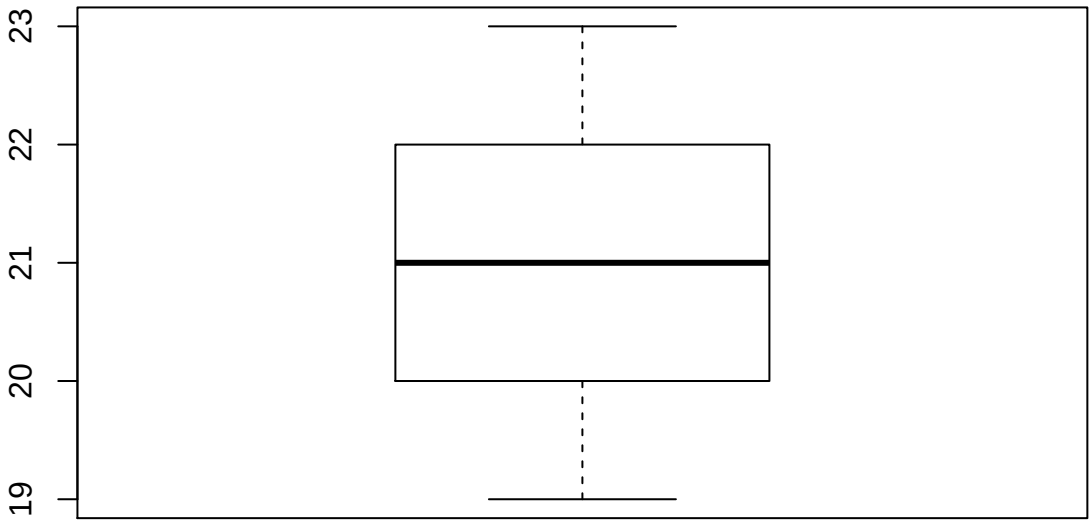

```r
# View a histogram of the absences variable
hist(students3$absences)
```



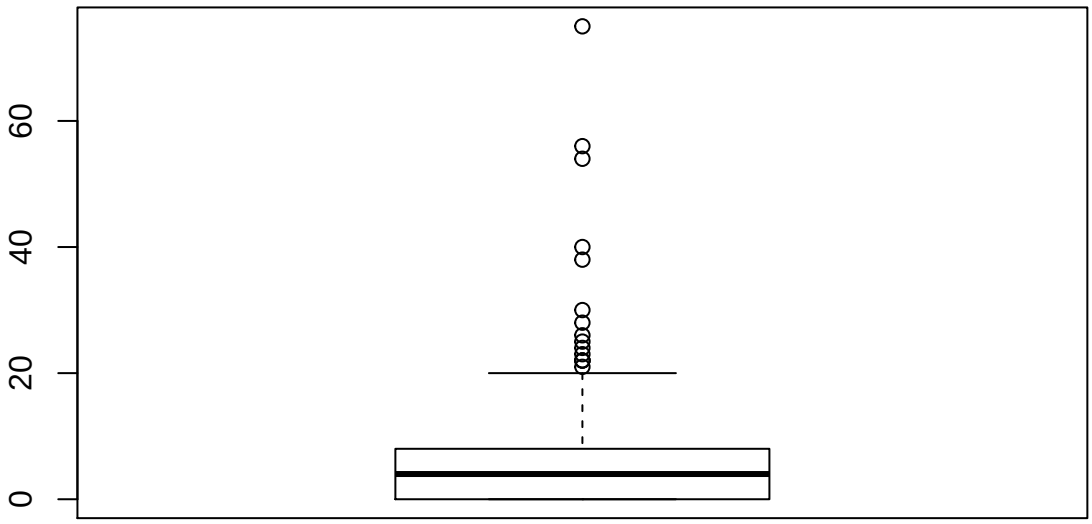

```
# View a histogram of absences, but force zeros to be bucketed to the right of zero
hist(students3$absences, right = FALSE)
```




```
# View a boxplot of age
boxplot(students3$age)
```




```
# View a boxplot of absences
boxplot(students3$absences)
```


0
20
40
60