1.-basics.R

jinyoungpark

2025-01-21

```
##### 1. R basics ############################################################
  tmp = c(1, 2, 3, 4)
  tmp <- c(1, 2, 3, 4)
  tmp == c(1, 2, 3, 4)
```

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

```
  tmp != c(1, 2, 3, 4)
```

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

```
  1+1
```

```
## [1] 2
```

```
  tmp*4
```

```
## [1]  4  8 12 16
```

```
  tmp2 = c("1","2","3","4")
  #tmp2*4 # does this work?
  as.numeric(tmp2)*4
```

```
## [1]  4  8 12 16
```

```
  tmp3 = data.frame(v1 = c(1,2,3,4), v2 = c(5,6,7,8))
  tmp3*4
```

```
##   v1 v2
## 1  4 20
## 2  8 24
## 3 12 28
## 4 16 32
```

```
  tmp3$v1 + tmp3$v2
```

```
## [1]  6  8 10 12
```

```r
#?mean # help
mean(tmp)
```

```
## [1] 2.5
```

```r
mean(tmp3)
```

```
## Warning in mean.default(tmp3): argument is not numeric or logical: returning NA
```

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

```r
mean(tmp3[,1])
```

```
## [1] 2.5
```

```r
mean(tmp3[,2])
```

```
## [1] 6.5
```

```r
apply(tmp3, 2, mean)
```

```
##  v1  v2
## 2.5 6.5
```

```r
apply(tmp3, 1, mean)
```

```
## [1] 3 4 5 6
```

```r
colMeans(tmp3)
```

```
##  v1  v2
## 2.5 6.5
```

```r
rowMeans(tmp3)
```

```
## [1] 3 4 5 6
```

```r
class(tmp3)
```

```
## [1] "data.frame"
```

```r
str(tmp3)
```

```
## 'data.frame':    4 obs. of  2 variables:
##  $ v1: num  1 2 3 4
##  $ v2: num  5 6 7 8
```

```r
  class(tmp3$v1)
```

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

```r
  class(tmp3$v2)
```

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

```r
  sapply(tmp3, class)
```

```
##        v1        v2
## "numeric" "numeric"
```

```r
  sapply(tmp3, is.numeric)
```

```
##   v1   v2
## TRUE TRUE
```

```r
  sapply(tmp3, is.na)
```

```
##         v1    v2
## [1,] FALSE FALSE
## [2,] FALSE FALSE
## [3,] FALSE FALSE
## [4,] FALSE FALSE
```

```r
  sapply(tmp3, is.factor)
```

```
##    v1    v2
## FALSE FALSE
```

```r
##### 1-2. examples ############################################################

  library(psych)

  #### descriptive stats ####
  data(sat.act)
  names(sat.act)
```

```
## [1] "gender"    "education" "age"       "ACT"       "SATV"      "SATQ"
```

```r
  head(sat.act)
```

```
##       gender education age ACT SATV SATQ
## 29442      2         3  19  24  500  500
## 29457      2         3  23  35  600  500
## 29498      2         3  20  21  480  470
## 29503      1         4  27  26  550  520
## 29504      1         2  33  31  600  550
## 29518      1         5  26  28  640  640
```

```
summary(sat.act)
```

```
##      gender         education          age              ACT
##  Min.   :1.000   Min.   :0.000   Min.   :13.00   Min.   : 3.00
##  1st Qu.:1.000   1st Qu.:3.000   1st Qu.:19.00   1st Qu.:25.00
##  Median :2.000   Median :3.000   Median :22.00   Median :29.00
##  Mean   :1.647   Mean   :3.164   Mean   :25.59   Mean   :28.55
##  3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:29.00   3rd Qu.:32.00
##  Max.   :2.000   Max.   :5.000   Max.   :65.00   Max.   :36.00
##
##       SATV            SATQ
##  Min.   :200.0   Min.   :200.0
##  1st Qu.:550.0   1st Qu.:530.0
##  Median :620.0   Median :620.0
##  Mean   :612.2   Mean   :610.2
##  3rd Qu.:700.0   3rd Qu.:700.0
##  Max.   :800.0   Max.   :800.0
##                  NA's   :13
```

```
psych::describe(sat.act)
```

```
##           vars   n   mean     sd median trimmed    mad min max range  skew
## gender       1 700   1.65   0.48      2    1.68   0.00   1   2     1 -0.61
## education    2 700   3.16   1.43      3    3.31   1.48   0   5     5 -0.68
## age          3 700  25.59   9.50     22   23.86   5.93  13  65    52  1.64
## ACT          4 700  28.55   4.82     29   28.84   4.45   3  36    33 -0.66
## SATV         5 700 612.23 112.90    620  619.45 118.61 200 800   600 -0.64
## SATQ         6 687 610.22 115.64    620  617.25 118.61 200 800   600 -0.59
##           kurtosis   se
## gender       -1.62 0.02
## education    -0.07 0.05
## age           2.42 0.36
## ACT           0.53 0.18
## SATV          0.33 4.27
## SATQ         -0.02 4.41
```

```
describeBy(sat.act, "gender")
```

```
##
##  Descriptive statistics by group
## gender: 1
##           vars   n   mean     sd median trimmed    mad min max range  skew
## gender       1 247   1.00   0.00      1    1.00   0.00   1   1     0   NaN
## education    2 247   3.00   1.54      3    3.12   1.48   0   5     5 -0.54
## age          3 247  25.86   9.74     22   24.23   5.93  14  58    44  1.43
## ACT          4 247  28.79   5.06     30   29.23   4.45   3  36    33 -1.06
## SATV         5 247 615.11 114.16    630  622.07 118.61 200 800   600 -0.63
## SATQ         6 245 635.87 116.02    660  645.53  94.89 300 800   500 -0.72
##           kurtosis   se
## gender         NaN 0.00
## education    -0.60 0.10
```

```
## age           1.43 0.62
## ACT           1.89 0.32
## SATV          0.13 7.26
## SATQ         -0.12 7.41
## ------------------------------------------------------------
## gender: 2
##           vars   n   mean     sd median trimmed    mad min max range  skew
## gender       1 453   2.00   0.00      2    2.00   0.00   2   2     0   NaN
## education    2 453   3.26   1.35      3    3.40   1.48   0   5     5 -0.74
## age          3 453  25.45   9.37     22   23.70   5.93  13  65    52  1.77
## ACT          4 453  28.42   4.69     29   28.63   4.45  15  36    21 -0.39
## SATV         5 453 610.66 112.31    620  617.91 103.78 200 800   600 -0.65
## SATQ         6 442 596.00 113.07    600  602.21 133.43 200 800   600 -0.58
##           kurtosis   se
## gender         NaN 0.00
## education     0.27 0.06
## age           3.03 0.44
## ACT          -0.42 0.22
## SATV          0.42 5.28
## SATQ          0.13 5.38
```

```
dim(sat.act) #700 6
```

```
## [1] 700   6
```

```
#### create summary variables ####
sat.act$sum.sat = apply(sat.act[,c("SATV", "SATQ")], 1, sum)
head(sat.act)
```

```
##       gender education age ACT SATV SATQ sum.sat
## 29442      2         3  19  24  500  500    1000
## 29457      2         3  23  35  600  500    1100
## 29498      2         3  20  21  480  470     950
## 29503      1         4  27  26  550  520    1070
## 29504      1         2  33  31  600  550    1150
## 29518      1         5  26  28  640  640    1280
```

```
#### correlations ####
tmp.cor = sat.act[1:10,c("ACT", "SATV", "SATQ")] # subsetting data
corr.test(tmp.cor)
```

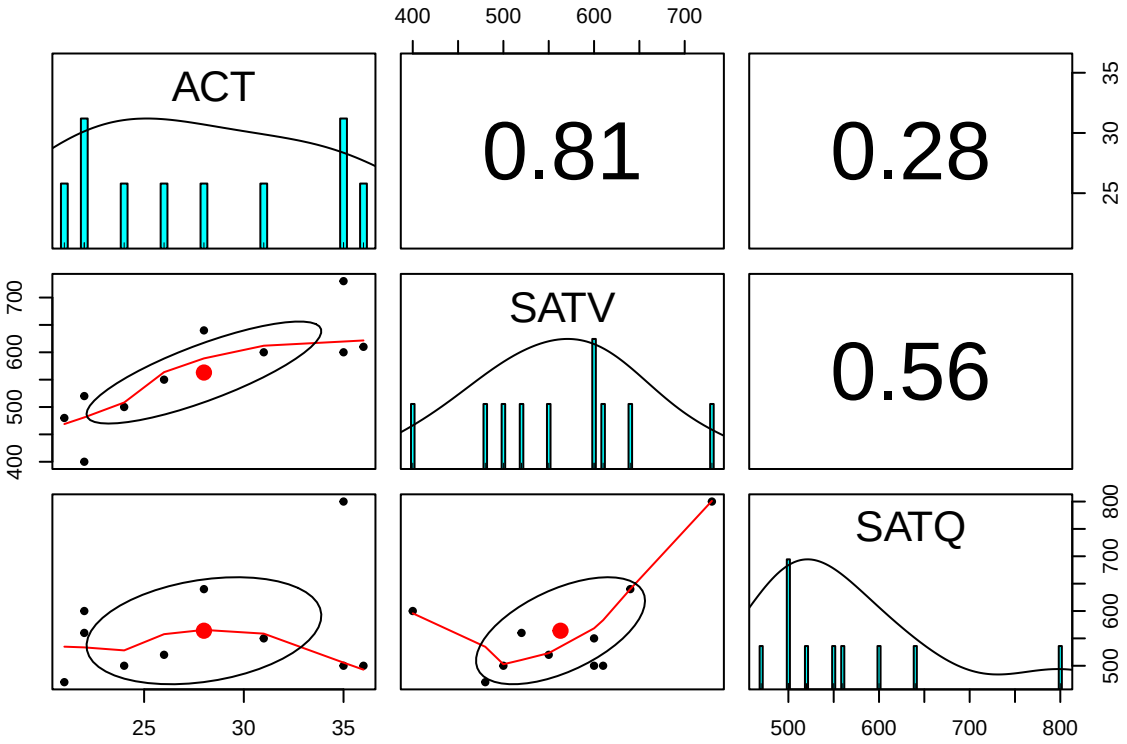
```
## Call:corr.test(x = tmp.cor)
## Correlation matrix
##       ACT SATV SATQ
## ACT  1.00 0.81 0.28
## SATV 0.81 1.00 0.56
## SATQ 0.28 0.56 1.00
## Sample Size
## [1] 10
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##       ACT SATV SATQ
## ACT  0.00 0.01 0.43
```

```
## SATV 0.00 0.00 0.18
## SATQ 0.43 0.09 0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

```
  lowerCor(tmp.cor)
```

```
##      ACT  SATV SATQ
## ACT  1.00
## SATV 0.81 1.00
## SATQ 0.28 0.56 1.00
```

```
  pairs.panels(tmp.cor)
```




```
##### 1-3. practice ############################################################
```

```
  library(haven); library(dplyr); library(magrittr); library(ggplot2)
```

```
##
## 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: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
  dat = read_sas("teeth.sas7bdat")
  dat$Sex %<>% as.factor()
  dat$Ethnicity %<>% as.factor()

  dat %>% psych::describe()
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf
```

```
##            vars  n   mean    sd median trimmed   mad min  max range  skew
## ID            1 66 133.94 21.48  131.5  132.94 23.72 101  176    75  0.38
## Visit         2 66   1.15  0.36    1.0    1.07  0.00   1    2     1  1.90
## VisitDate     3 64    NaN    NA     NA     NaN    NA Inf -Inf  -Inf    NA
## Age           4 65  38.98 17.81   34.0   37.32 16.31  18   85    67  0.73
## Sex*          5 66   1.38  0.49    1.0    1.35  0.00   1    2     1  0.49
## Race          6 65   2.03  1.94    1.0    1.57  0.00   1    9     8  2.57
## Ethnicity*    7 64   2.06  0.64    2.0    2.00  0.00   1    4     3  1.38
## Nteeth        8 66  23.79  8.60   27.5   25.17  5.93   0   32    32 -1.34
##            kurtosis   se
## ID            -0.98 2.64
## Visit          1.63 0.04
## VisitDate        NA   NA
## Age           -0.52 2.21
## Sex*          -1.79 0.06
## Race           6.11 0.24
## Ethnicity*     3.32 0.08
## Nteeth         1.04 1.06
```

```
  dat %>% psych::describeBy(dat$Sex)
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf
```

```
##
##  Descriptive statistics by group
## group: 1
##            vars  n   mean    sd median trimmed   mad min  max range  skew
## ID            1 41 136.12 21.26    134  135.06 22.24 103  176    73  0.35
## Visit         2 41   1.15  0.36      1    1.06  0.00   1    2     1  1.93
```

```
## VisitDate    3  0    NaN    NA     NA     NaN    NA Inf -Inf  -Inf    NA
## Age          4 41  37.98 17.03     34   36.36 16.31  18   85    67  0.73
## Sex          5 41   1.00  0.00      1    1.00  0.00   1    1     0   NaN
## Race         6 41   1.56  0.81      1    1.39  0.00   1    4     3  1.47
## Ethnicity    7 41   2.02  0.57      2    2.00  0.00   1    4     3  1.59
## Nteeth       8 41  23.51  8.80     27   24.91  5.93   0   32    32 -1.26
##           kurtosis   se
## ID           -0.92 3.32
## Visit         1.76 0.06
## VisitDate       NA   NA
## Age          -0.31 2.66
## Sex            NaN 0.00
## Race          1.71 0.13
## Ethnicity     5.29 0.09
## Nteeth        0.79 1.37
## ------------------------------------------------------------
## group: 2
##           vars  n   mean    sd median trimmed   mad min  max range  skew
## ID           1 25 130.36 21.79    123  129.52 23.72 101  169    68  0.44
## Visit        2 25   1.16  0.37      1    1.10  0.00   1    2     1  1.74
## VisitDate    3  0    NaN    NA     NA     NaN    NA Inf -Inf  -Inf    NA
## Age          4 24  40.71 19.32     34   39.30 14.83  18   80    62  0.64
## Sex          5 25   2.00  0.00      2    2.00  0.00   2    2     0   NaN
## Race         6 24   2.83  2.88      1    2.40  0.00   1    9     8  1.33
## Ethnicity    7 23   2.13  0.76      2    2.05  0.00   1    4     3  1.00
## Nteeth       8 25  24.24  8.43     28   25.57  5.93   0   32    32 -1.39
##           kurtosis   se
## ID           -1.21 4.36
## Visit         1.09 0.07
## VisitDate       NA   NA
## Age          -1.04 3.94
## Sex            NaN 0.00
## Race          0.13 0.59
## Ethnicity     1.03 0.16
## Nteeth        1.16 1.69
```

```
dat %>% summary
```

```
##        ID            Visit         VisitDate                Age        Sex
##  Min.   :101.0   Min.   :1.000   Min.   :2013-08-01   Min.   :18.00   1:41
##  1st Qu.:116.2   1st Qu.:1.000   1st Qu.:2013-10-14   1st Qu.:25.00   2:25
##  Median :131.5   Median :1.000   Median :2013-11-08   Median :34.00
##  Mean   :133.9   Mean   :1.152   Mean   :2013-11-02   Mean   :38.98
##  3rd Qu.:147.8   3rd Qu.:1.000   3rd Qu.:2013-11-30   3rd Qu.:54.00
##  Max.   :176.0   Max.   :2.000   Max.   :2013-12-31   Max.   :85.00
##                                  NA's   :2            NA's   :1
##       Race       Ethnicity     Nteeth
##  Min.   :1.000   1   : 7   Min.   : 0.00
##  1st Qu.:1.000   2   :50   1st Qu.:20.00
##  Median :1.000   8   : 3   Median :27.50
##  Mean   :2.031   9   : 4   Mean   :23.79
##  3rd Qu.:2.000   NA's: 2   3rd Qu.:30.00
##  Max.   :9.000             Max.   :32.00
##  NA's   :1
```

```r
table(dat$Sex)
```

```
##
##  1  2
## 41 25
```

```r
table(dat$Nteeth)
```

```
##
##  0  2  5 12 14 15 16 18 19 20 22 23 24 25 26 27 28 29 30 31 32
##  3  1  1  2  1  4  2  1  1  2  2  3  3  3  1  3 10  3  8  2 10
```

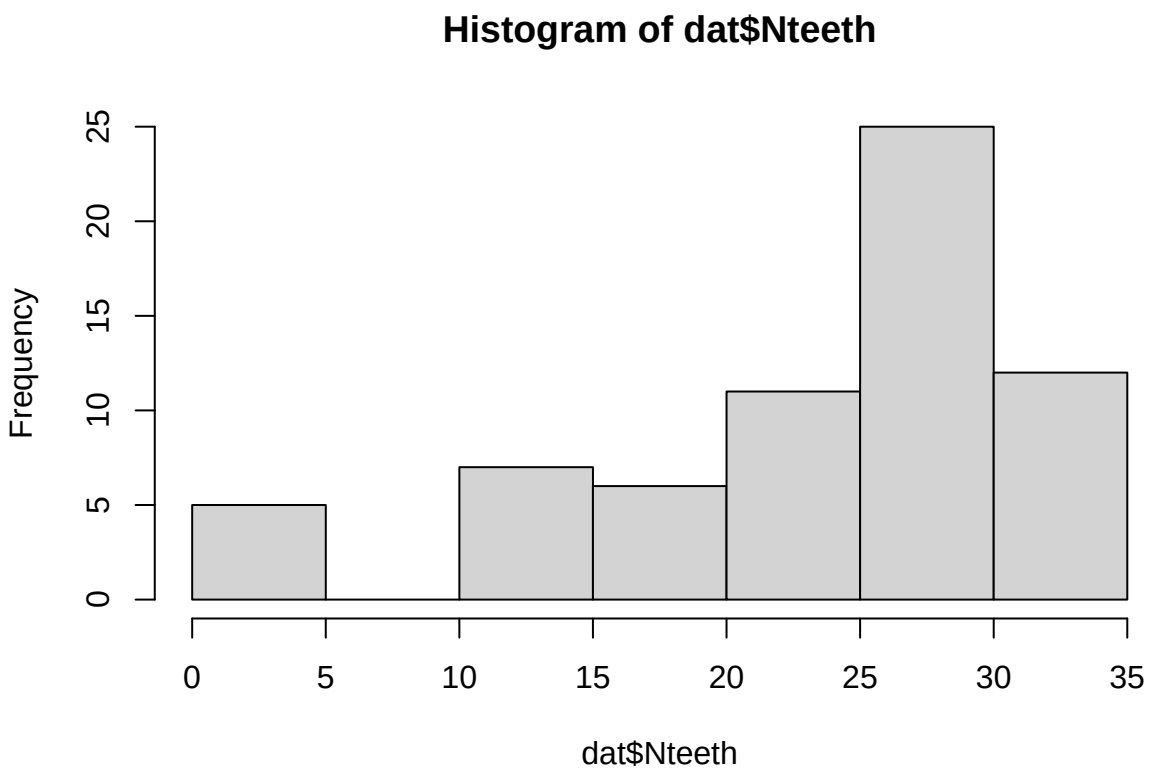
```r
table(dat$Sex, dat$Nteeth)
```

```
##
##     0 2 5 12 14 15 16 18 19 20 22 23 24 25 26 27 28 29 30 31 32
##   1 2 1 0  2  0  3  1  1  1  1  1  2  2  1  1  3  6  0  5  2  6
##   2 1 0 1  0  1  1  1  0  0  1  1  1  1  2  0  0  4  3  3  0  4
```

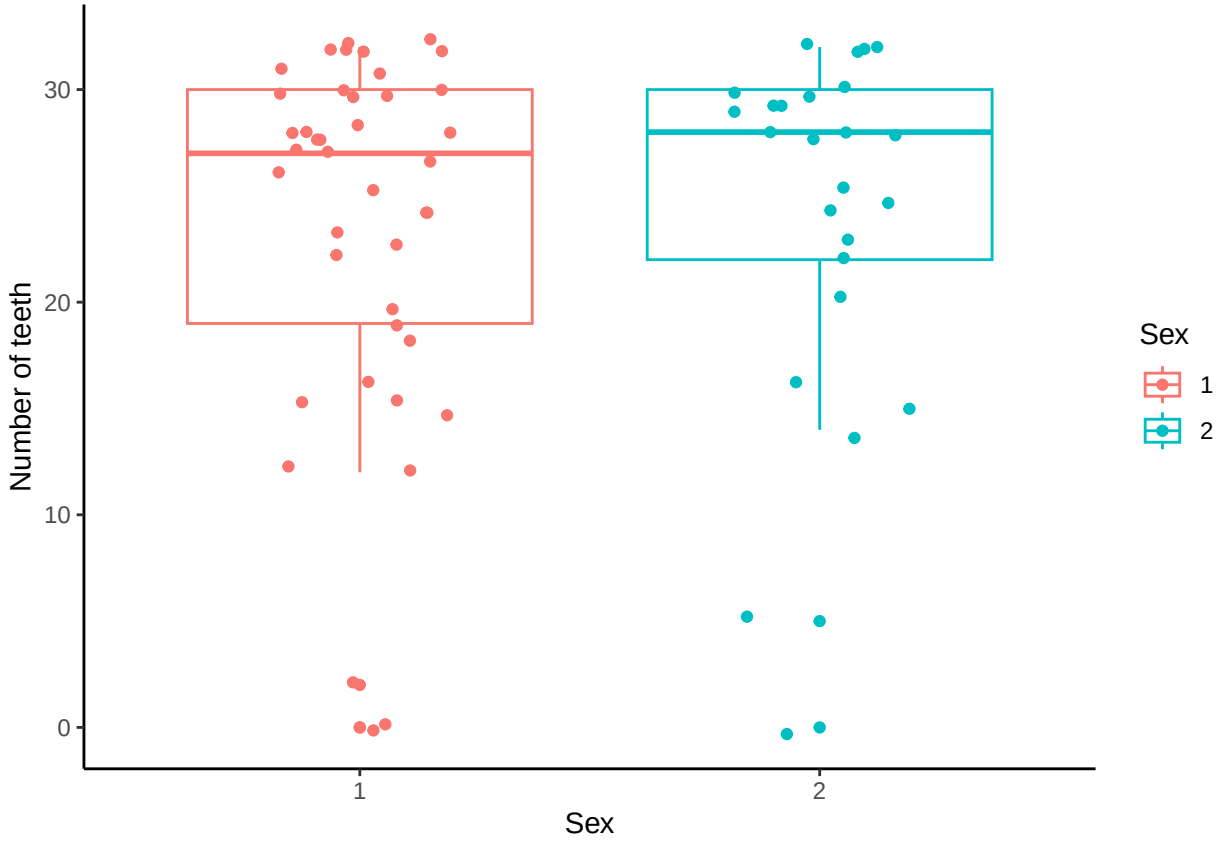
```r
dat %>% group_by(Sex) %>%
  summarise(mean.teeth = mean(Nteeth), n = n(), sd = sd(Nteeth))
```

```
## # A tibble: 2 x 4
##   Sex   mean.teeth     n    sd
##   <fct>      <dbl> <int> <dbl>
## 1 1           23.5    41  8.80
## 2 2           24.2    25  8.43
```

```r
hist(dat$Nteeth) # some graphics
```

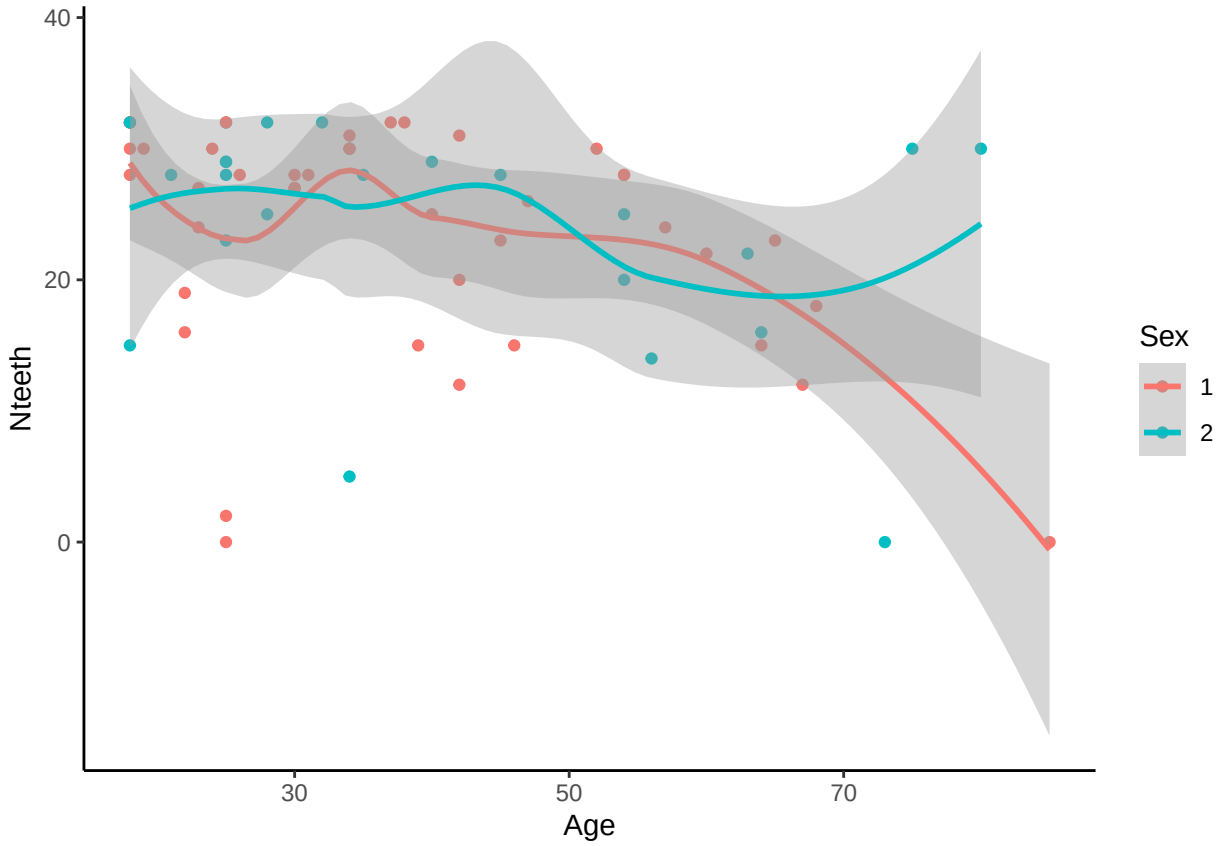


```r
dat %>%
  ggplot(aes(Sex, Nteeth, group = Sex, color = Sex)) +
  geom_boxplot() +
  geom_jitter(width = 0.2) +
  ylab("Number of teeth")  +
  theme_classic()
```




```r
dat %>%
  ggplot(aes(Age, Nteeth, group = Sex, color = Sex)) +
  geom_point() +
  geom_smooth(method = loess) +
  theme_classic()
```

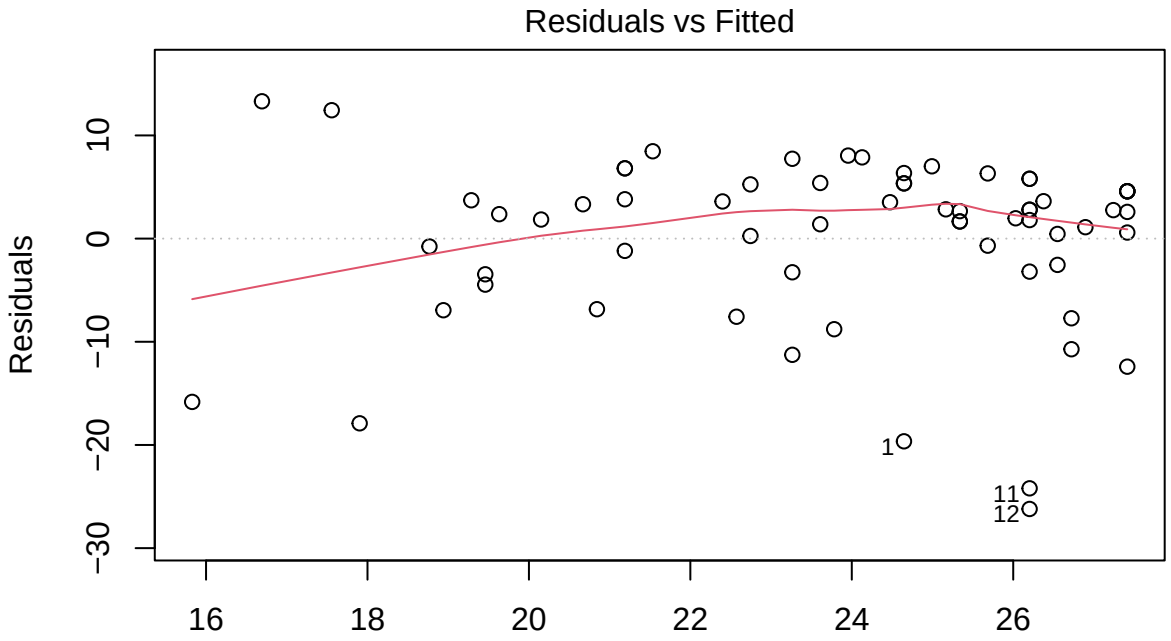
```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```

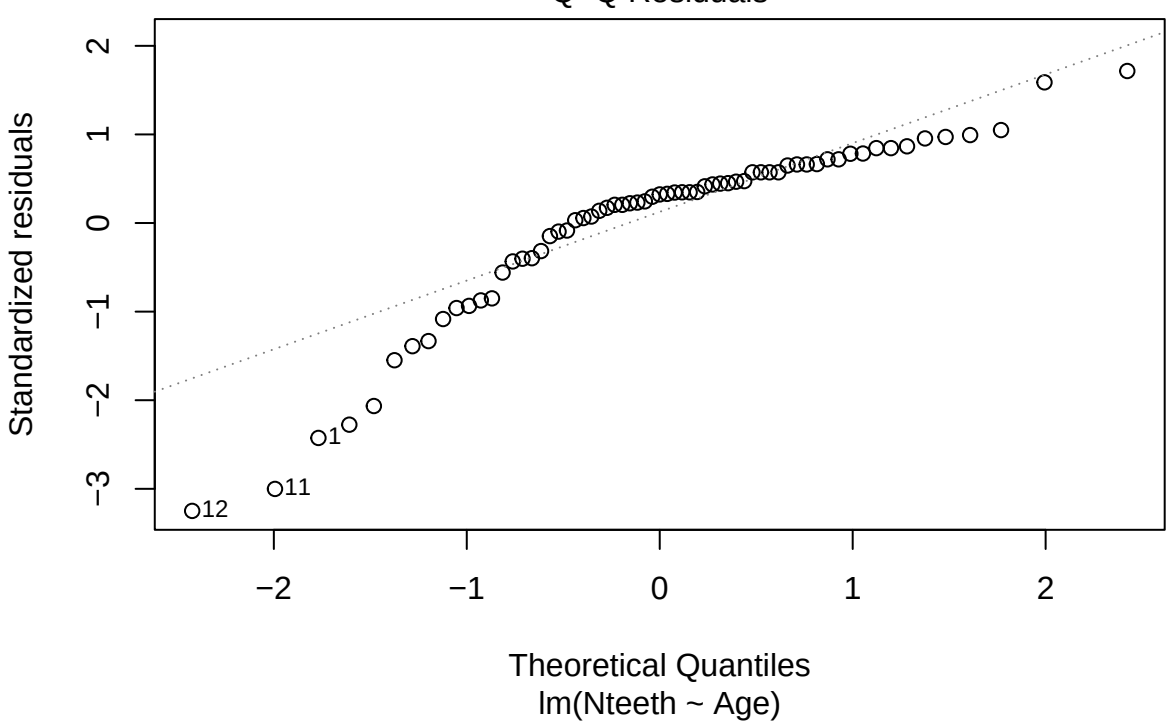


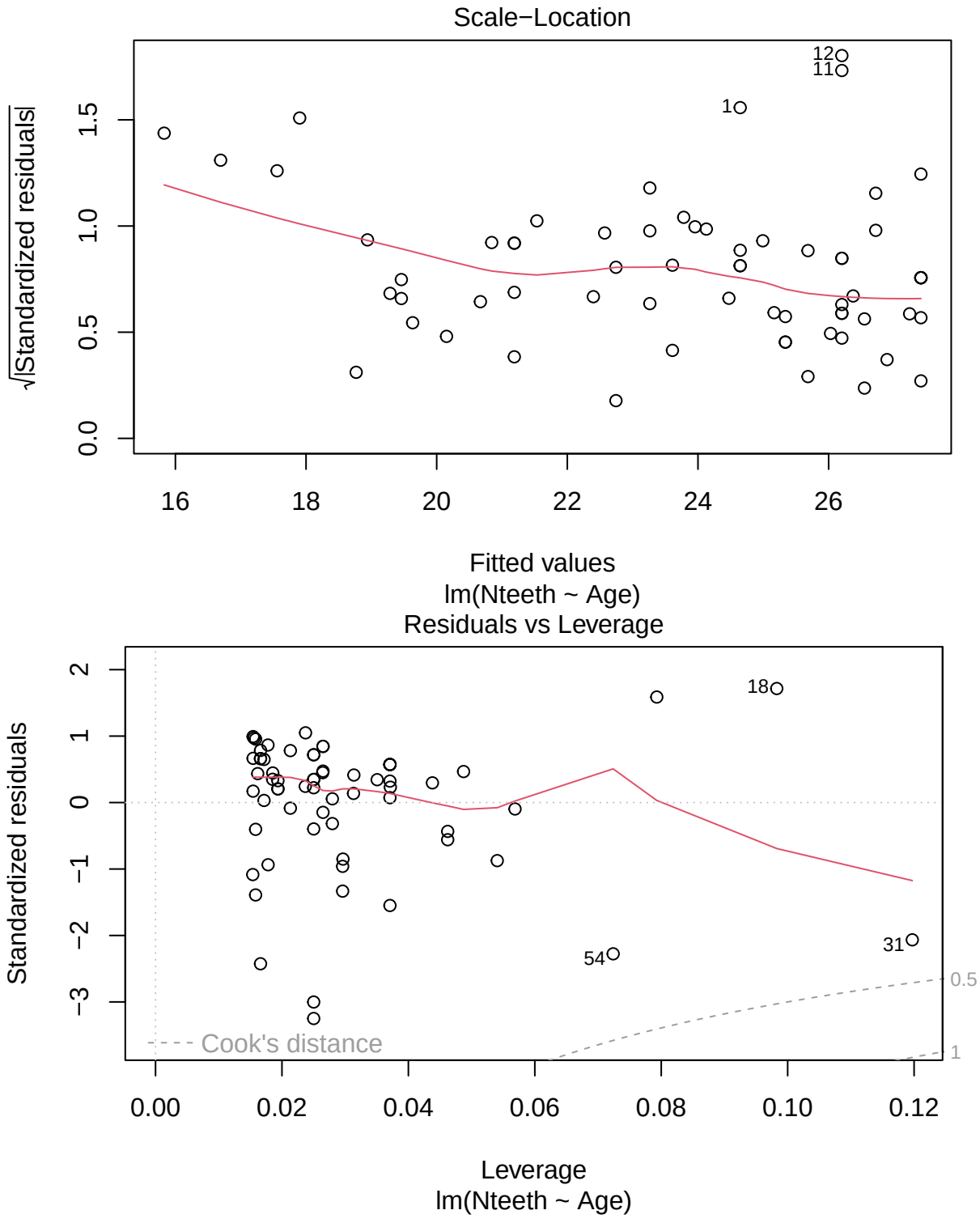

```
  fit = lm(Nteeth ~ Age, data = dat)
  fit %>% summary
```

```
##
## Call:
## lm(formula = Nteeth ~ Age, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.203  -3.203   2.587   5.255  13.307
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.52546    2.45381  12.440  < 2e-16 ***
## Age         -0.17291    0.05733  -3.016  0.00369 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.167 on 63 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1262, Adjusted R-squared:  0.1123
## F-statistic: 9.097 on 1 and 63 DF,  p-value: 0.003689
```

```
  plot(fit)
```


Residuals vs Fitted
Residuals
10
0
−10
−20
−30
16
18
20
22
24
26
1
11
12
Fitted values
lm(Nteeth ~ Age)


Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−3
−2
−1
0
1
2
1
11
12
Theoretical Quantiles
lm(Nteeth ~ Age)


Scale−Location
√|Standardized residuals|
Fitted values
lm(Nteeth ~ Age)
Residuals vs Leverage
Standardized residuals
Cook's distance
0.5
1
Leverage
lm(Nteeth ~ Age)