Parkinson data - I

Giovanni Saraceno

Contents

```
library(tidyverse)
library(dplyr)
library(ggplot2)
```

We consider the data set related to the study in Tsanas et al. (2009). Data can be found at https:// archive.ics.uci.edu/ ml/ datasets/ Parkinsons+Telemonitoring.

We start loading the data.

```
dat <- read.csv('parkinsons_updrs.data', he = TRUE, sep = ',')
str(dat)
```

```
## 'data.frame':    5875 obs. of  22 variables:
##  $ subject.     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age          : int  72 72 72 72 72 72 72 72 72 72 ...
##  $ sex          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ test_time    : num  5.64 12.67 19.68 25.65 33.64 ...
##  $ motor_UPDRS  : num  28.2 28.4 28.7 28.9 29.2 ...
##  $ total_UPDRS  : num  34.4 34.9 35.4 35.8 36.4 ...
##  $ Jitter...    : num  0.00662 0.003 0.00481 0.00528 0.00335 0.00353 0.00422 0.00476 0.00432 0.00496
##  $ Jitter.Abs.  : num  3.38e-05 1.68e-05 2.46e-05 2.66e-05 2.01e-05 ...
##  $ Jitter.RAP   : num  0.00401 0.00132 0.00205 0.00191 0.00093 0.00119 0.00212 0.00226 0.00156 0.002
##  $ Jitter.PPQ5  : num  0.00317 0.0015 0.00208 0.00264 0.0013 0.00159 0.00221 0.00259 0.00207 0.00253
##  $ Jitter.DDP   : num  0.01204 0.00395 0.00616 0.00573 0.00278 ...
##  $ Shimmer      : num  0.0256 0.0202 0.0168 0.0231 0.017 ...
##  $ Shimmer.dB.  : num  0.23 0.179 0.181 0.327 0.176 0.214 0.445 0.212 0.371 0.31 ...
##  $ Shimmer.APQ3 : num  0.01438 0.00994 0.00734 0.01106 0.00679 ...
##  $ Shimmer.APQ5 : num  0.01309 0.01072 0.00844 0.01265 0.00929 ...
##  $ Shimmer.APQ11: num  0.0166 0.0169 0.0146 0.0196 0.0182 ...
##  $ Shimmer.DDA  : num  0.0431 0.0298 0.022 0.0332 0.0204 ...
##  $ NHR          : num  0.0143 0.0111 0.0202 0.0278 0.0116 ...
##  $ HNR          : num  21.6 27.2 23 24.4 26.1 ...
##  $ RPDE         : num  0.419 0.435 0.462 0.487 0.472 ...
##  $ DFA          : num  0.548 0.565 0.544 0.578 0.561 ...
##  $ PPE          : num  0.16 0.108 0.21 0.333 0.194 ...
```

Description of the variables

- subject# - Integer that uniquely identifies each subject
- `age` - Subject age
- `sex` - Subject gender '0' - male, '1' - female
- `test_time` - Time since recruitment into the trial. The integer part is the number of days since recruitment.

- `motor_UPDRS` - Clinician's motor UPDRS score, linearly interpolated
- `total_UPDRS` - Clinician's total UPDRS score, linearly interpolated
- `Jitter(%)`, `Jitter(Abs)`, `Jitter:RAP`, `Jitter:PPQ5`, `Jitter:DDP` - Several measures of variation in fundamental frequency
- `Shimmer`, `Shimmer(dB)`, `Shimmer:APQ3`, `Shimmer:APQ5`, `Shimmer:APQ11`, `Shimmer:DDA` - Several measures of variation in amplitude
- `NHR`, `HNR` - Two measures of ratio of noise to tonal components in the voice
- `RPDE` - A nonlinear dynamical complexity measure
- `DFA` - Signal fractal scaling exponent
- `PPE` - A nonlinear measure of fundamental frequency variation

```
str(dat)
```

```
## 'data.frame':    5875 obs. of  22 variables:
##  $ subject.     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age          : int  72 72 72 72 72 72 72 72 72 72 ...
##  $ sex          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ test_time    : num  5.64 12.67 19.68 25.65 33.64 ...
##  $ motor_UPDRS  : num  28.2 28.4 28.7 28.9 29.2 ...
##  $ total_UPDRS  : num  34.4 34.9 35.4 35.8 36.4 ...
##  $ Jitter...    : num  0.00662 0.003 0.00481 0.00528 0.00335 0.00353 0.00422 0.00476 0.00432 0.00496
##  $ Jitter.Abs.  : num  3.38e-05 1.68e-05 2.46e-05 2.66e-05 2.01e-05 ...
##  $ Jitter.RAP   : num  0.00401 0.00132 0.00205 0.00191 0.00093 0.00119 0.00212 0.00226 0.00156 0.002
##  $ Jitter.PPQ5  : num  0.00317 0.0015 0.00208 0.00264 0.0013 0.00159 0.00221 0.00259 0.00207 0.00253
##  $ Jitter.DDP   : num  0.01204 0.00395 0.00616 0.00573 0.00278 ...
##  $ Shimmer      : num  0.0256 0.0202 0.0168 0.0231 0.017 ...
##  $ Shimmer.dB.  : num  0.23 0.179 0.181 0.327 0.176 0.214 0.445 0.212 0.371 0.31 ...
##  $ Shimmer.APQ3 : num  0.01438 0.00994 0.00734 0.01106 0.00679 ...
##  $ Shimmer.APQ5 : num  0.01309 0.01072 0.00844 0.01265 0.00929 ...
##  $ Shimmer.APQ11: num  0.0166 0.0169 0.0146 0.0196 0.0182 ...
##  $ Shimmer.DDA  : num  0.0431 0.0298 0.022 0.0332 0.0204 ...
##  $ NHR          : num  0.0143 0.0111 0.0202 0.0278 0.0116 ...
##  $ HNR          : num  21.6 27.2 23 24.4 26.1 ...
##  $ RPDE         : num  0.419 0.435 0.462 0.487 0.472 ...
##  $ DFA          : num  0.548 0.565 0.544 0.578 0.561 ...
##  $ PPE          : num  0.16 0.108 0.21 0.333 0.194 ...
```

```
dat$subject. <- as.factor(dat$subject.)
dat$sex <- ifelse(dat$sex == 1, 'F', 'M')
dat$sex <- as.factor(dat$sex)
```

Data cleaning

Descriptive analysis Let's have a look at a brief descriptive summary of all the variables

```
summary(dat)
```

```
##     subject.         age       sex        test_time        motor_UPDRS
##  29     : 168   Min.   :36.0   F:1867   Min.   : -4.263   Min.   : 5.038
##  35     : 165   1st Qu.:58.0   M:4008   1st Qu.: 46.847   1st Qu.:15.000
##  41     : 165   Median :65.0            Median : 91.523   Median :20.871
##  7      : 161   Mean   :64.8            Mean   : 92.864   Mean   :21.296
##  34     : 161   3rd Qu.:72.0            3rd Qu.:138.445   3rd Qu.:27.596
##  5      : 156   Max.   :85.0            Max.   :215.490   Max.   :39.511
##  (Other):4899
```

```
##   total_UPDRS      Jitter...         Jitter.Abs.          Jitter.RAP
##  Min.   : 7.00   Min.   :0.000830   Min.   :2.250e-06   Min.   :0.000330
##  1st Qu.:21.37   1st Qu.:0.003580   1st Qu.:2.244e-05   1st Qu.:0.001580
##  Median :27.58   Median :0.004900   Median :3.453e-05   Median :0.002250
##  Mean   :29.02   Mean   :0.006154   Mean   :4.403e-05   Mean   :0.002987
##  3rd Qu.:36.40   3rd Qu.:0.006800   3rd Qu.:5.333e-05   3rd Qu.:0.003290
##  Max.   :54.99   Max.   :0.099990   Max.   :4.456e-04   Max.   :0.057540
##
##   Jitter.PPQ5         Jitter.DDP          Shimmer         Shimmer.dB.
##  Min.   :0.000430   Min.   :0.000980   Min.   :0.00306   Min.   :0.026
##  1st Qu.:0.001820   1st Qu.:0.004730   1st Qu.:0.01912   1st Qu.:0.175
##  Median :0.002490   Median :0.006750   Median :0.02751   Median :0.253
##  Mean   :0.003277   Mean   :0.008962   Mean   :0.03404   Mean   :0.311
##  3rd Qu.:0.003460   3rd Qu.:0.009870   3rd Qu.:0.03975   3rd Qu.:0.365
##  Max.   :0.069560   Max.   :0.172630   Max.   :0.26863   Max.   :2.107
##
##   Shimmer.APQ3      Shimmer.APQ5     Shimmer.APQ11      Shimmer.DDA
##  Min.   :0.00161   Min.   :0.00194   Min.   :0.00249   Min.   :0.00484
##  1st Qu.:0.00928   1st Qu.:0.01079   1st Qu.:0.01566   1st Qu.:0.02783
##  Median :0.01370   Median :0.01594   Median :0.02271   Median :0.04111
##  Mean   :0.01716   Mean   :0.02014   Mean   :0.02748   Mean   :0.05147
##  3rd Qu.:0.02057   3rd Qu.:0.02375   3rd Qu.:0.03272   3rd Qu.:0.06173
##  Max.   :0.16267   Max.   :0.16702   Max.   :0.27546   Max.   :0.48802
##
##       NHR                HNR              RPDE             DFA
##  Min.   :0.000286   Min.   : 1.659   Min.   :0.1510   Min.   :0.5140
##  1st Qu.:0.010955   1st Qu.:19.406   1st Qu.:0.4698   1st Qu.:0.5962
##  Median :0.018448   Median :21.920   Median :0.5423   Median :0.6436
##  Mean   :0.032120   Mean   :21.680   Mean   :0.5415   Mean   :0.6532
##  3rd Qu.:0.031463   3rd Qu.:24.444   3rd Qu.:0.6140   3rd Qu.:0.7113
##  Max.   :0.748260   Max.   :37.875   Max.   :0.9661   Max.   :0.8656
##
##       PPE
##  Min.   :0.02198
##  1st Qu.:0.15634
##  Median :0.20550
##  Mean   :0.21959
##  3rd Qu.:0.26449
##  Max.   :0.73173
##
```

We want to check whether NAs are present.

```
colSums(is.na(dat))
```

```
##      subject.           age           sex     test_time   motor_UPDRS
##             0             0             0             0             0
##   total_UPDRS     Jitter...   Jitter.Abs.    Jitter.RAP   Jitter.PPQ5
##             0             0             0             0             0
##    Jitter.DDP       Shimmer   Shimmer.dB.  Shimmer.APQ3  Shimmer.APQ5
##             0             0             0             0             0
## Shimmer.APQ11   Shimmer.DDA           NHR           HNR          RPDE
##             0             0             0             0             0
##           DFA           PPE
##             0             0
```

We can also check the variance of each variable, in order to look for 'almost' constant variable (variance very close to zero).

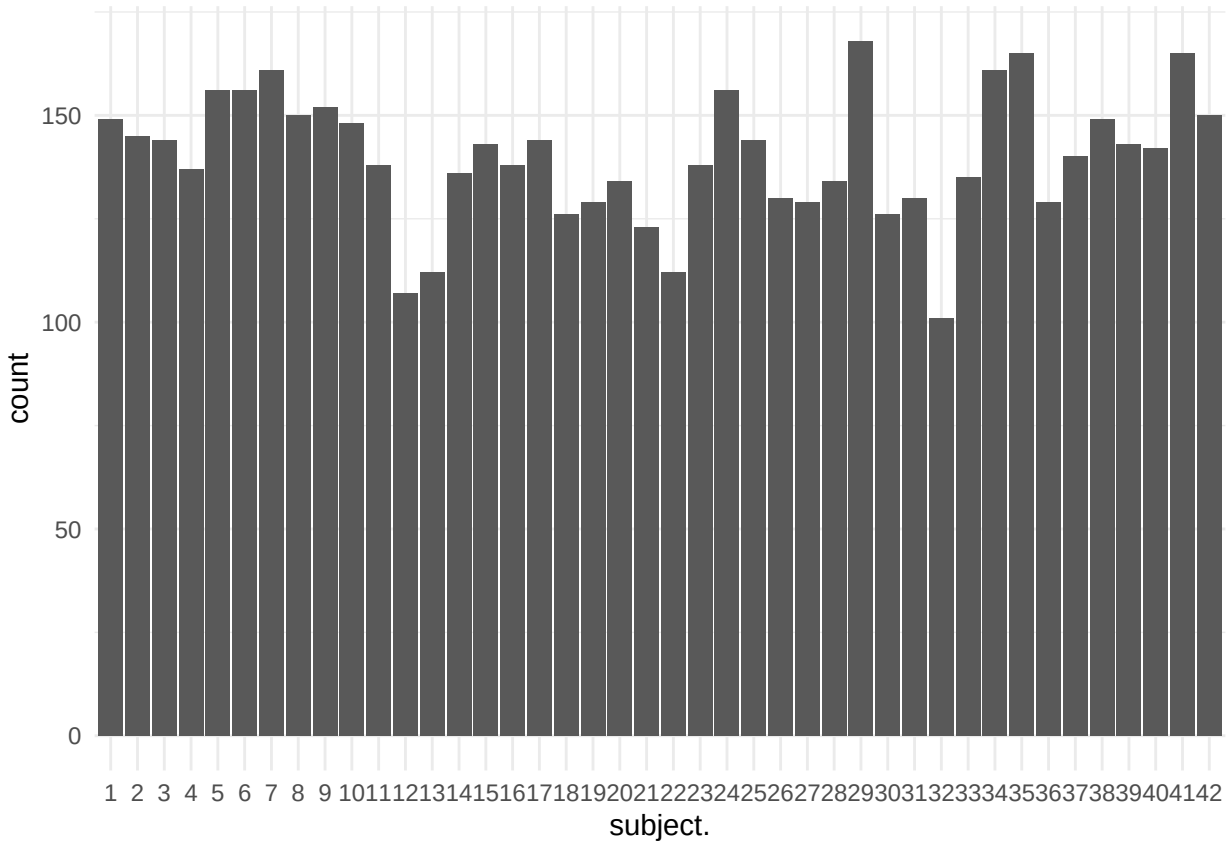
```
apply(dat, 2, function(x) var(x))
```

```
## Warning in var(x): NA introdotti per coercizione
```

```
##      subject.           age           sex     test_time   motor_UPDRS
##  1.530733e+02  7.781928e+01            NA  2.856432e+03  6.608522e+01
##   total_UPDRS     Jitter...   Jitter.Abs.    Jitter.RAP   Jitter.PPQ5
##  1.144961e+02  3.163194e-05  1.294802e-09  9.758245e-06  1.392436e-05
##    Jitter.DDP       Shimmer   Shimmer.dB.  Shimmer.APQ3  Shimmer.APQ5
##  8.782483e-05  6.674554e-04  5.301675e-02  1.752213e-04  2.776867e-04
## Shimmer.APQ11   Shimmer.DDA           NHR           HNR          RPDE
##  3.994460e-04  1.576995e-03  3.563171e-03  1.841351e+01  1.019813e-02
##           DFA           PPE
##  5.027093e-03  8.371974e-03
```

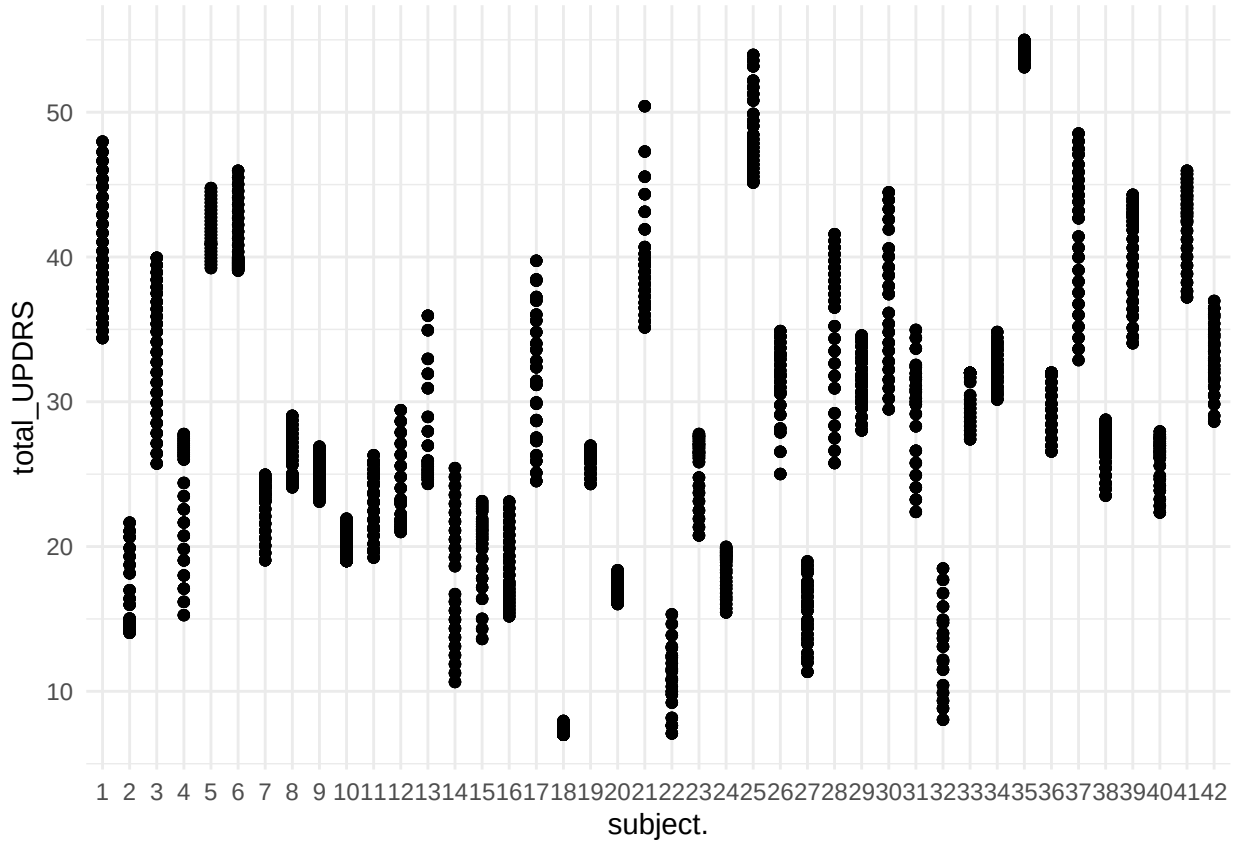
We can now investigates the variables in details. We can start with the categorical variables and in relation to `total_UPDRS` since it is our variable of interest.
We start with `subject`

```
ggplot(dat, aes(x=subject.)) +
  geom_bar() +
  theme_minimal()
```



We have different number of measurements fo the subjects.

```
ggplot(dat, aes(x = subject., y = total_UPDRS)) +
  geom_point() +
  theme_minimal()
```

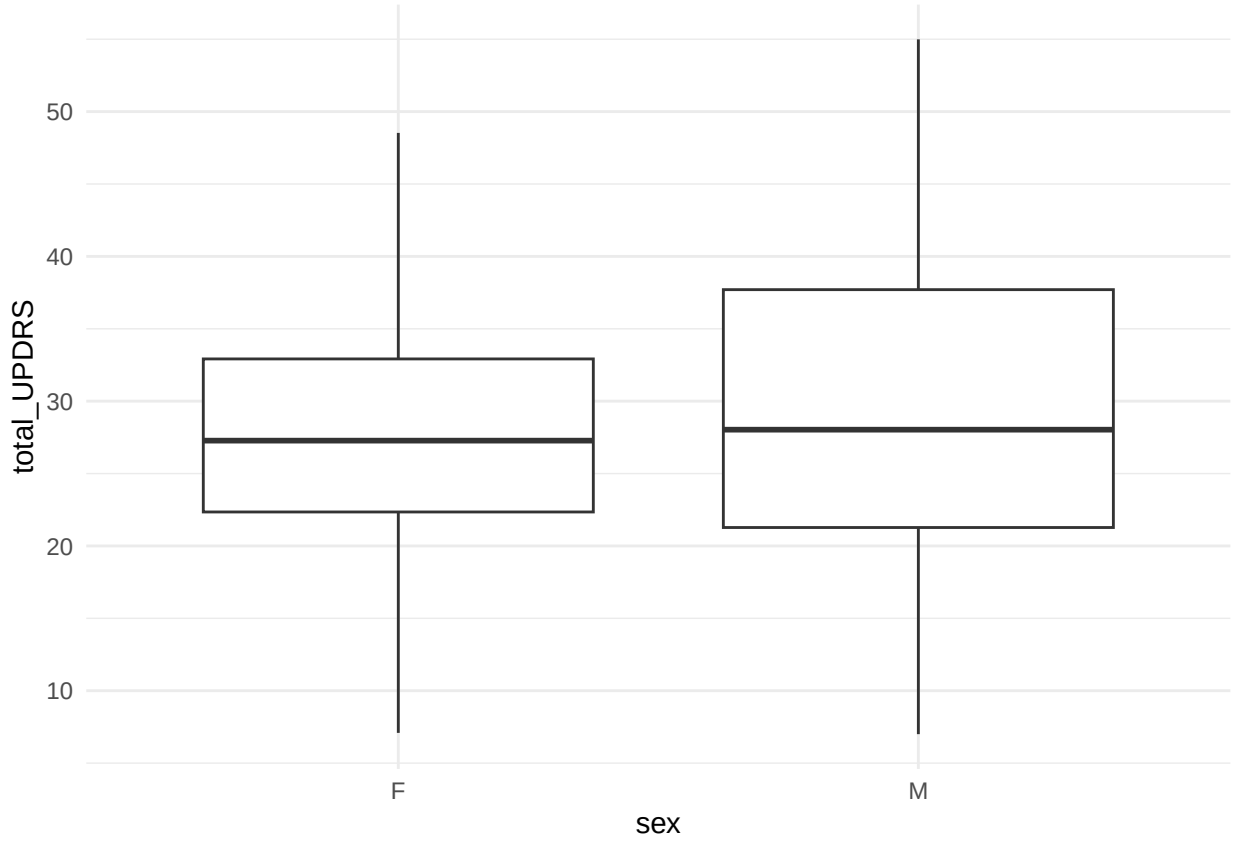


It is useful also to explore the sex variable

```
dat %>% group_by(sex) %>% summarise(n = n())
```

```
## # A tibble: 2 x 2
##   sex       n
##   <fct> <int>
## 1 F      1867
## 2 M      4008
```

```
ggplot(dat, aes(x = sex, y = total_UPDRS)) +
  geom_boxplot() +
  theme_minimal()
```

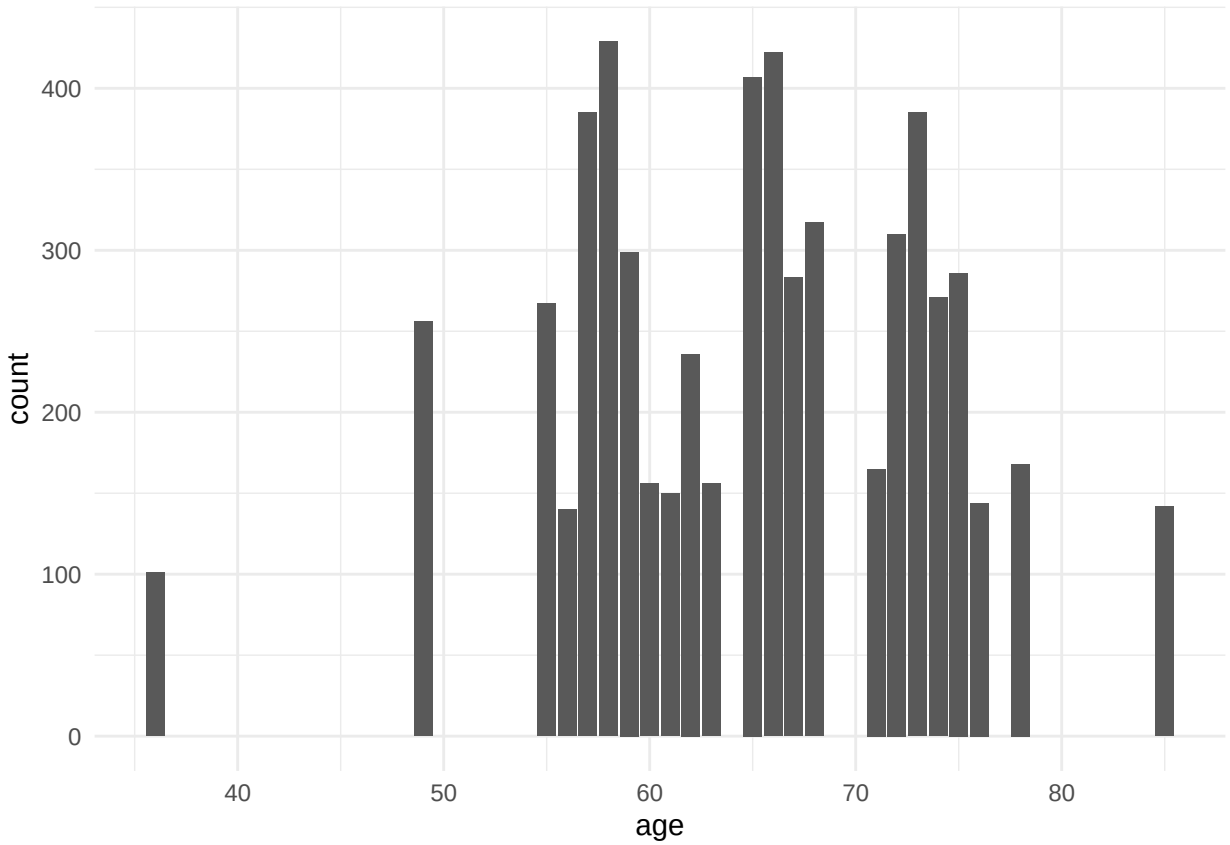


Now let's look at *age*

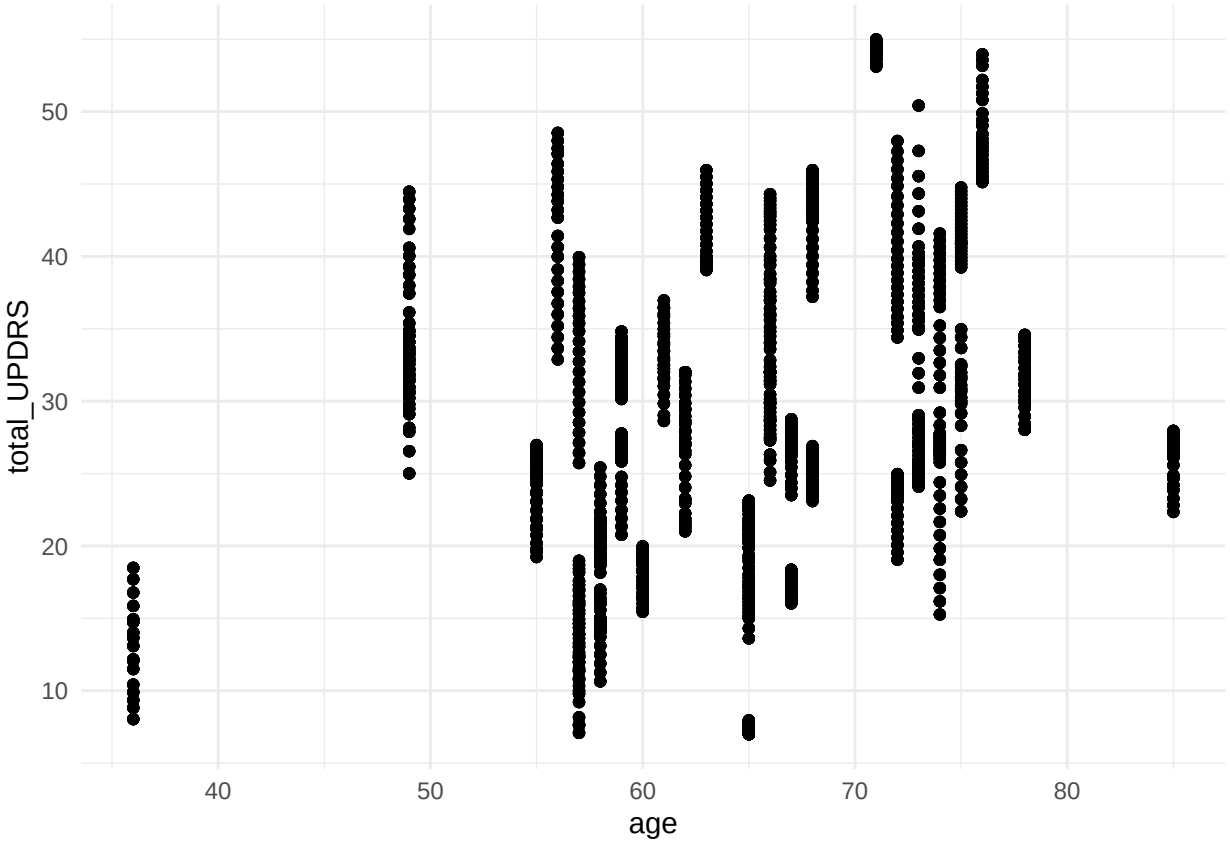
```
dat %>% group_by(age) %>% summarise(n = n())
```

```
## # A tibble: 23 x 2
##      age     n
##    <int> <int>
##  1    36   101
##  2    49   256
##  3    55   267
##  4    56   140
##  5    57   385
##  6    58   429
##  7    59   299
##  8    60   156
##  9    61   150
## 10    62   236
## # i 13 more rows
```

```
ggplot(dat, aes(x = age)) +
  geom_bar() +
  theme_minimal()
```

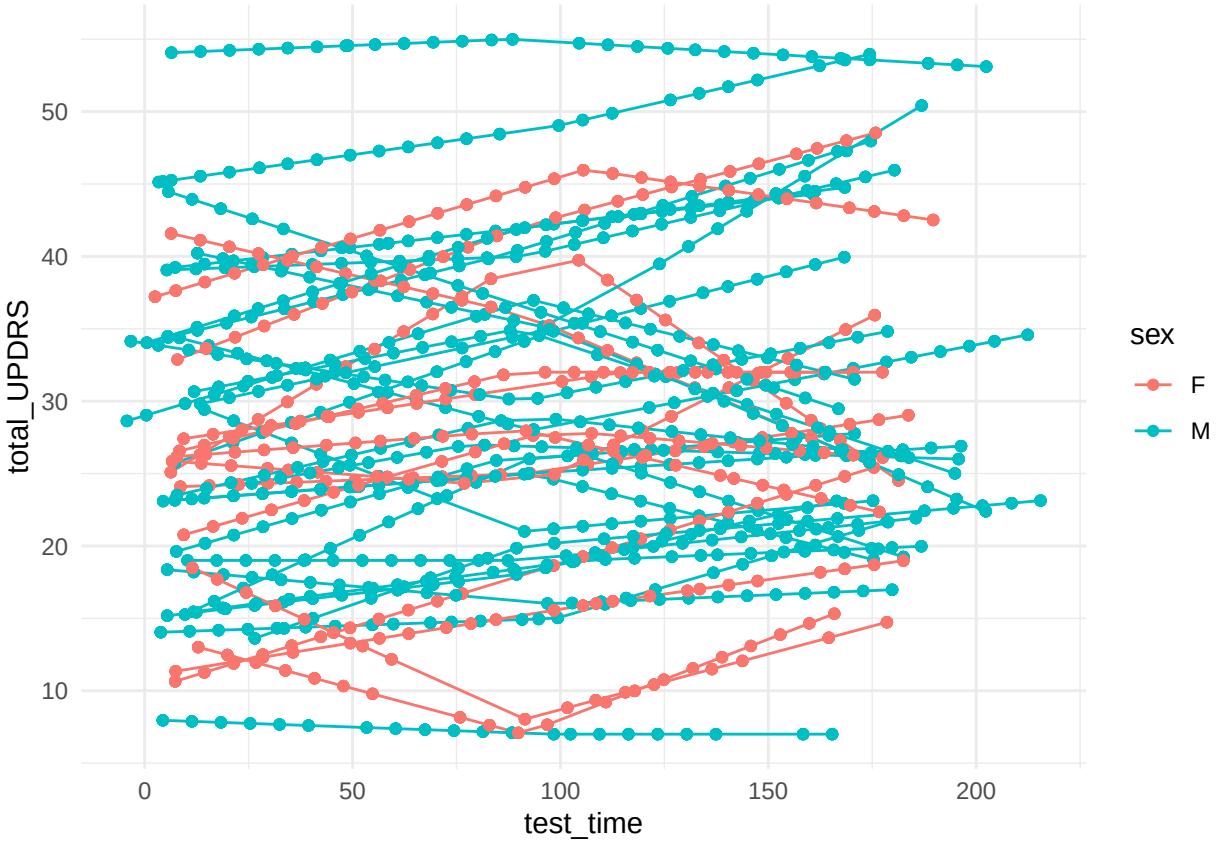



```
ggplot(dat, aes(x = age, y = total_UPDRS)) +
  geom_point() +
  theme_minimal()
```



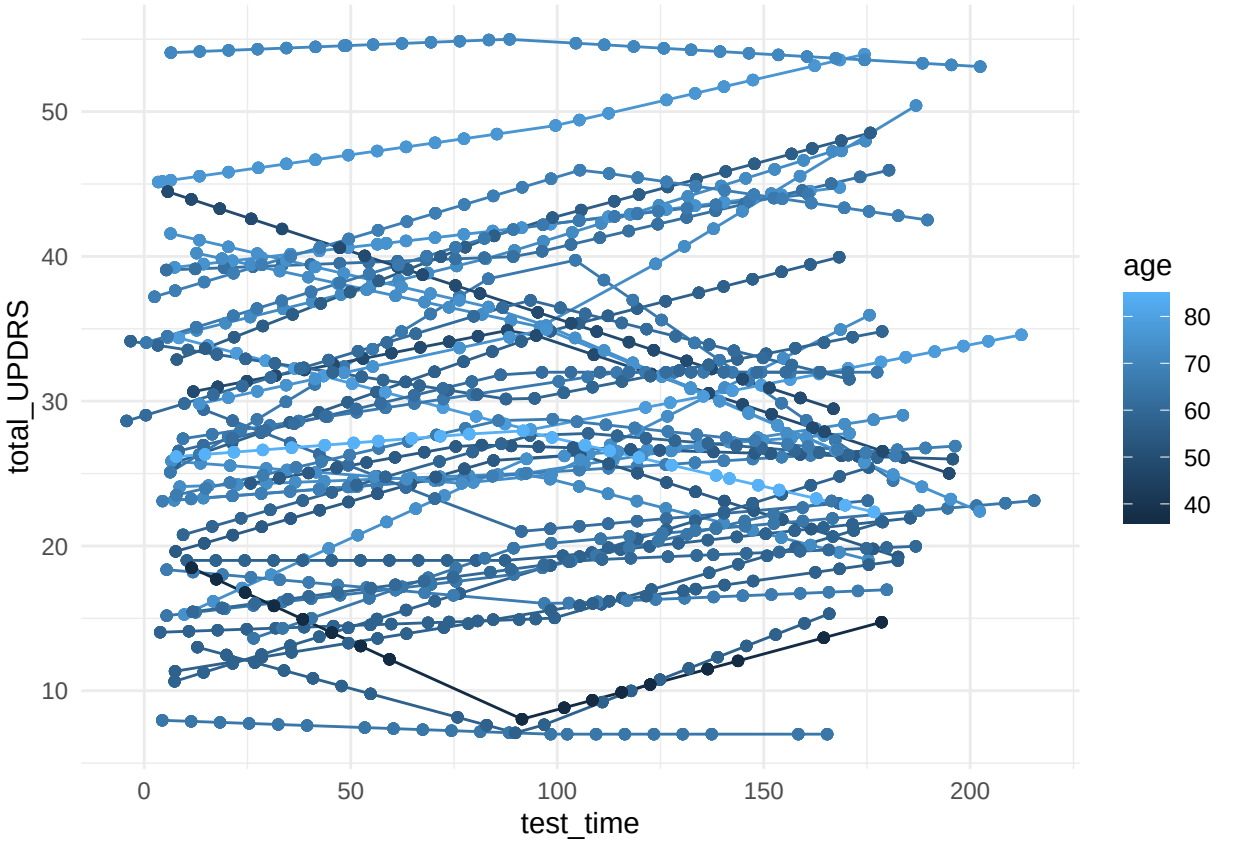
There is also the variable that indicates the time at which the data are collected for each subject. Let us look at the values of `total_UPDRS` in time, grouped by subject and colored by sex

```
ggplot(dat, aes(x = test_time, y = total_UPDRS, group = subject., col=sex)) +
  geom_line() +
  geom_point() +
  theme_minimal()
```



or by age

```
ggplot(dat, aes(x = test_time, y = total_UPDRS, group = subject., col=age)) +
  geom_line() +
  geom_point() +
  theme_minimal()
```



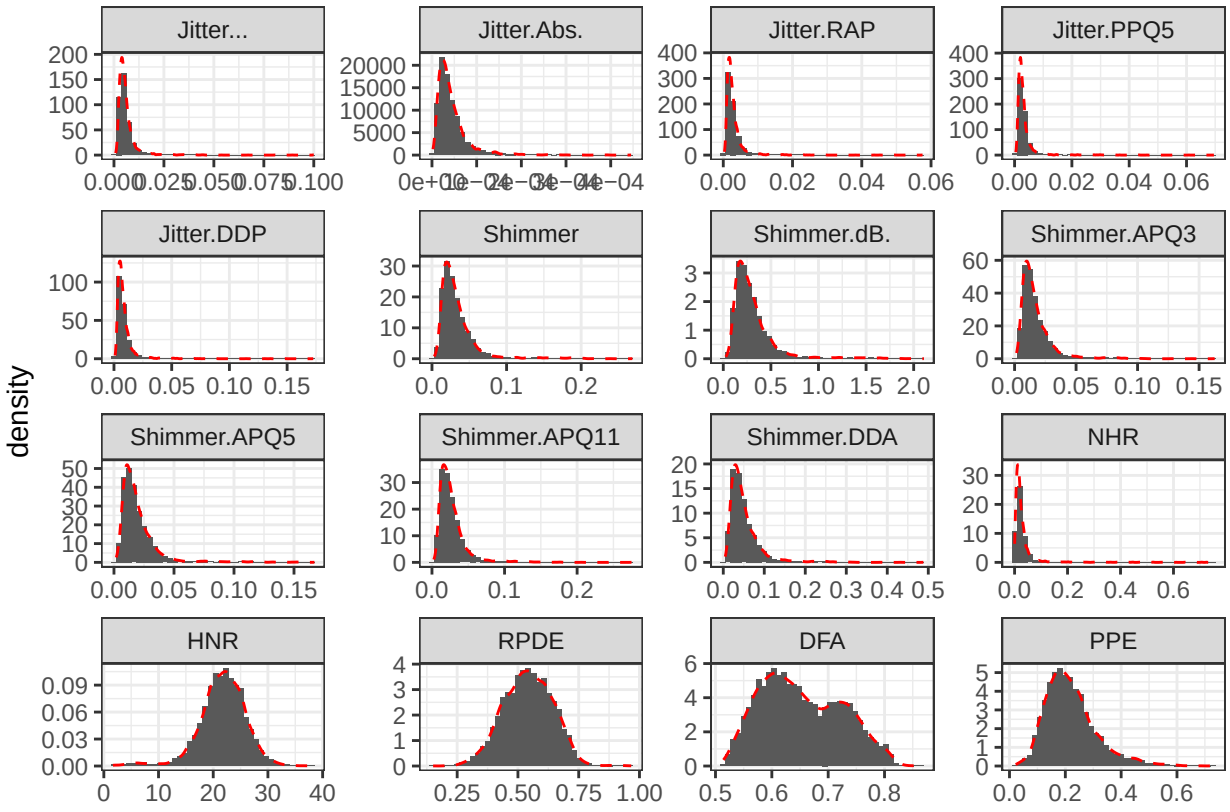
Consider now the covariates

```
library(reshape2)
```

```
## Warning: il pacchetto 'reshape2' è stato creato con R versione 4.3.2
```

```
dat_dens <- melt(dat[, 7:22])
ggplot(dat_dens, aes(x = value)) +
  geom_histogram(aes(y=..density..), bins=40) +
  geom_density(color="red", linetype="dashed") +
  facet_wrap(~ variable, ncol = 4, scales = 'free') +
  theme_bw() +
  xlab("")
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



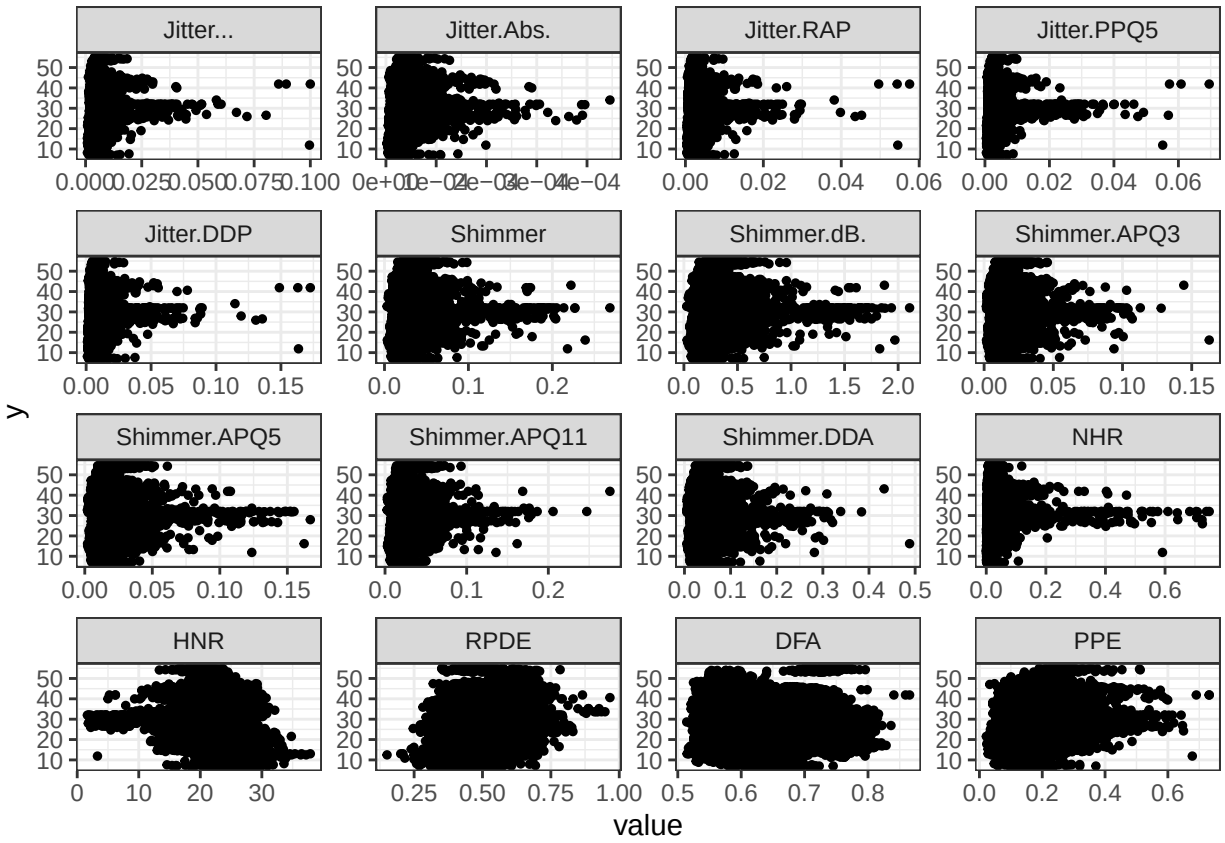
We can see that `HNR` and `RPDE` are bell-shaped, while `DFA` seems to be bi-modal.
We can continue computing the correlation of these variables with the `total_UPDRS`

```
apply(dat[, 7:22], 2, function(x) cor(dat$total_UPDRS, x, method = 'spearman'))
```

```
##     Jitter...   Jitter.Abs.    Jitter.RAP   Jitter.PPQ5    Jitter.DDP
##     0.1292373     0.1041782     0.1092140     0.1183471     0.1092514
##       Shimmer   Shimmer.dB.  Shimmer.APQ3  Shimmer.APQ5 Shimmer.APQ11
##     0.1375502     0.1399155     0.1199090     0.1249397     0.1611506
##   Shimmer.DDA           NHR           HNR          RPDE           DFA
##     0.1199116     0.1439723    -0.1622837     0.1499256    -0.1415379
##           PPE
##     0.1552364
```

and the corresponding scatter plots

```
dat_dens$y <- rep(dat$total_UPDRS, length(7:22))
ggplot(dat_dens, aes(x = value, y = y)) +
  geom_point(size = 1) +
  facet_wrap(~ variable, ncol = 4, scales = 'free') +
  theme_bw()
```

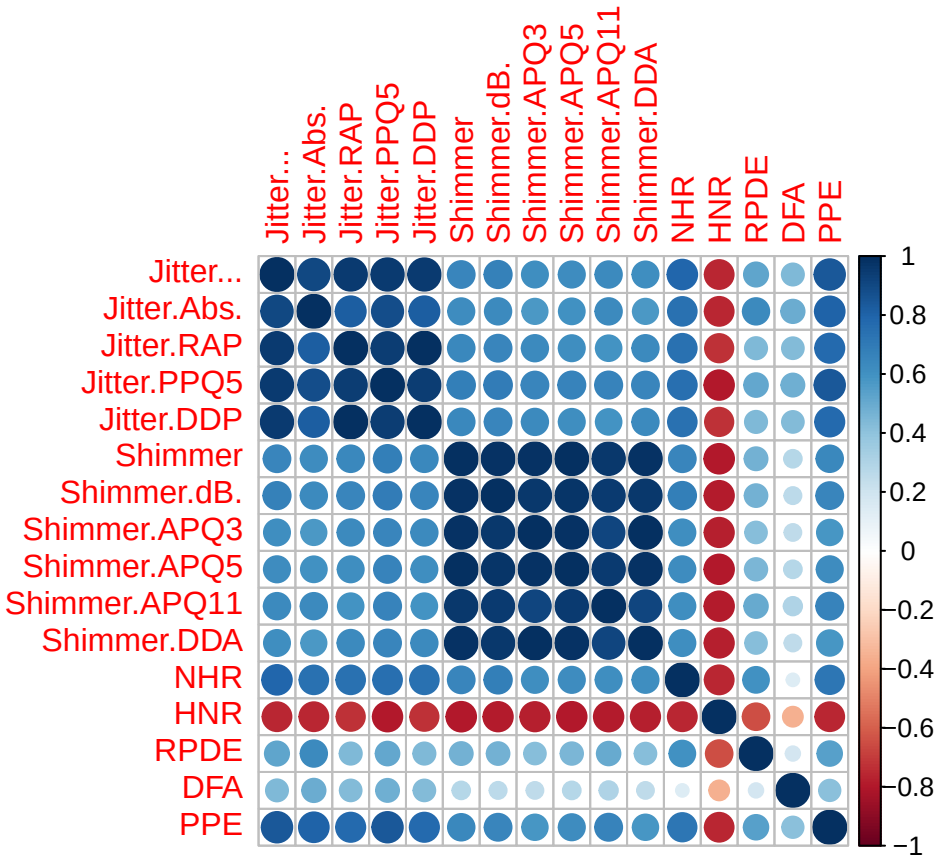


Finally we compute the correlations among the covariates

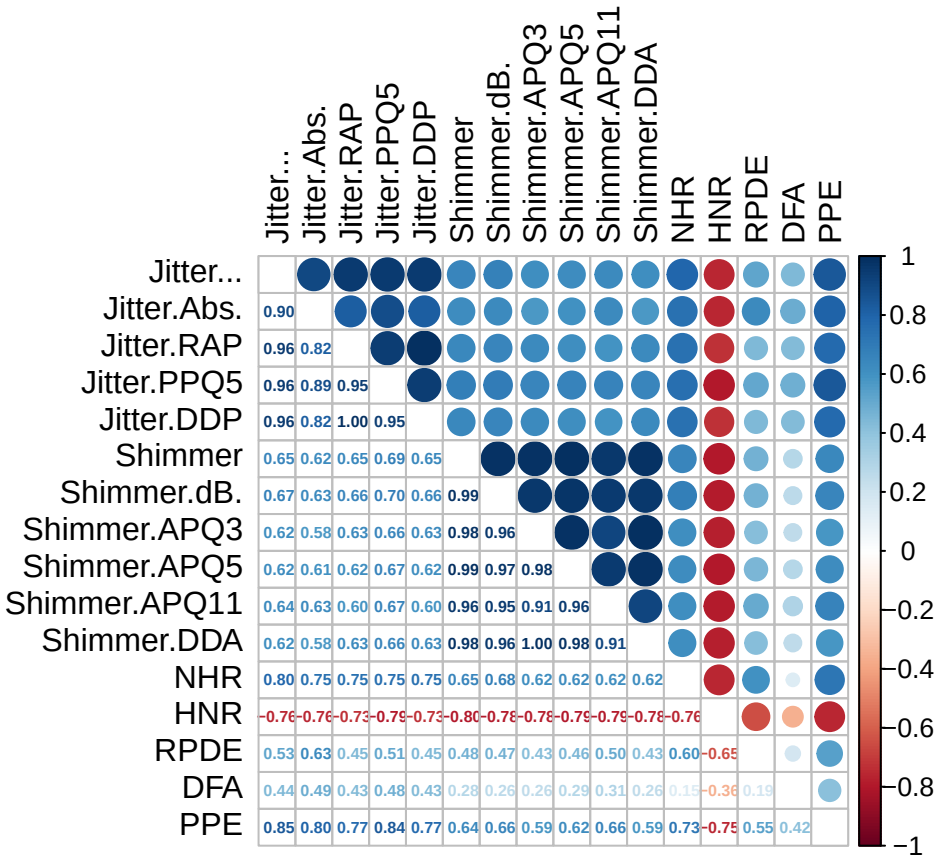
```
cor_cov <- cor(dat[, 7:22], method = 'spearman')
library(corrplot)
```

```
## corrplot 0.94 loaded
```

```
corrplot(cor_cov)
```




```
corrplot.mixed(cor_cov, tl.col = 'black', insig = 'blank', tl.pos = 'lt', number.cex = 0.5)
```



References Tsanas, Athanasios, Max Little, Patrick McSharry, and Lorraine Ramig. 2009. "Accurate Telemonitoring of Parkinson's Disease Progression by Non-Invasive Speech Tests." *Nature Precedings*, 1–1.