EDA_ALSPAC

2025-06-09

```
library(foreign)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Reading in ALSPAC data

```
alspac = read.spss("Ulven_B3777_13Jan22.sav",to.data.frame = TRUE,add.undeclared.levels = "no")
```

```
## re-encoding from CP1252
```

```
#str(alspac)
```

Checking data structure

```
str(alspac)
```

```
## 'data.frame':    15646 obs. of  58 variables:
##  $ cidB3777          : num  1 2 3 4 6 7 8 10 12 13 ...
##  $ qlet              : chr  "A" "A" "A" "A" ...
##  $ areases_quint_preg: num  NA NA 1 3 1 5 5 5 5 3 ...
##  $ no2_preg          : num  NA NA 22 26.9 29.6 26.5 27.9 27.4 33.3 27.9 ...
##  $ pm25_preg         : num  NA NA 14.5 13.5 13.3 13.4 13.4 13 14.5 13.8 ...
##  $ pm10_preg         : num  NA NA 27.8 29.8 NA ...
##  $ kz011b            : Factor w/ 4 levels "Consent withdrawn",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ kz021             : Factor w/ 4 levels "Consent withdrawn",..: 4 4 4 3 3 3 4 4 4 4 ...
##  $ kz030             : num  NA NA 3520 3570 3700 3340 2740 2900 3970 3600 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-10" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Not in core sample" "Triplet / quadruplet" "Consent withdrawn"
##  $ chol_cord         : num  NA NA NA 0.77 NA 1.46 NA 1.85 2.06 NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
```

```
##  $ trig_cord         : num  NA NA NA 0.41 NA 0.34 NA 0.7 0.63 NA ...
##   ..- attr(*, "value.labels")= Named chr [1:4] "-1" "-2" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:4] "Missing/Not assayed" "Out of detectable range (<0.1 or >10 mmo
##  $ HDL_cord          : num  NA NA NA 0.23 NA 0.35 NA 0.38 0.74 NA ...
##   ..- attr(*, "value.labels")= Named chr [1:5] "-1" "-2" "-8" "-11" ...
##   .. ..- attr(*, "names")= chr [1:5] "Missing/Not assayed" "Out of detectable range (<0.08 or >3.12 
##  $ LDL_cord          : num  NA NA NA 0.354 NA ...
##   ..- attr(*, "value.labels")= Named chr [1:4] "-1" "-8" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:4] "Missing/Not assayed" "Unresolvable" "Trip/Quad" "Consent withd
##  $ CRP_cord          : num  NA NA NA 0.06 NA 0 NA 0 0 NA ...
##   ..- attr(*, "value.labels")= Named chr [1:4] "-1" "-8" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:4] "Missing/Not assayed" "Unresolvable" "Trip/Quad" "Consent withd
##  $ Trig_CIF31        : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ Chol_CIF31        : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ LDL_CIF31         : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ HDL_CIF31         : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ Trig_CIF43        : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ Chol_CIF43        : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ LDL_CIF43         : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ HDL_CIF43         : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ Chol_BBS          : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ Trig_BBS          : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ HDL_BBS           : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ LDL_BBS           : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ CHOL_F7           : num  5.75 NA NA NA NA NA NA NA NA 4.37 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ TRIG_F7           : num  1.83 NA NA NA NA NA NA NA NA 1.37 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
```

```
##  $ HDL_F7            : num  1.26 NA NA NA NA NA NA NA NA 1.69 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ LDL_F7            : num  3.61 NA NA NA NA NA NA NA NA 2.06 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ CHOL_F9           : num  5.42 NA 4.09 NA 3.83 NA NA NA NA 3.98 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ trig_f9           : num  0.74 NA 1.61 NA 0.37 NA NA NA NA 0.91 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ HDL_f9            : num  1.47 NA 1.21 NA 1.67 NA NA NA NA 1.61 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ LDL_f9            : num  3.61 NA 2.14 NA 1.99 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ VLDL_f9           : num  0.338 NA 0.735 NA 0.169 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ CRP_f9            : num  0.28 NA 0.11 NA 0.22 NA NA NA NA 0.58 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ IL6_f9            : num  0.901 NA 1.139 NA 0.662 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ age_31m           : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-2" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing" "Did not attend" "Consent withdrawn"
##  $ age_43m           : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-2" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing" "Did not attend" "Consent withdrawn"
##  $ age_7y            : num  7.65 9.15 NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr "-9999"
##   .. ..- attr(*, "names")= chr "Consent withdrawn"
##  $ age_8y            : num  8.42 10.03 9.28 NA 8.76 ...
##   ..- attr(*, "value.labels")= Named chr "-9999"
##   .. ..- attr(*, "names")= chr "Consent withdrawn"
##  $ age_9y            : num  9.67 10.15 10.3 NA 9.88 ...
##   ..- attr(*, "value.labels")= Named chr "-9999"
##   .. ..- attr(*, "names")= chr "Consent withdrawn"
##  $ mz010a            : Factor w/ 4 levels "Consent withdrawn",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ mz028b            : num  NA NA 33 23 29 31 18 25 25 29 ...
##   ..- attr(*, "value.labels")= Named chr [1:8] "44" "15" "-1" "-2" ...
##   .. ..- attr(*, "names")= chr [1:8] ">43" "< 16" "Ma's DOB NK" "Miscarried" ...
##  $ a525              : Factor w/ 9 levels "Consent withdrawn",..: NA NA 7 NA 6 7 3 3 7 7 ...
##  $ b032              : num  NA NA 1 NA 1 0 0 0 0 0 ...
##   ..- attr(*, "value.labels")= Named chr [1:4] "-1" "-2" "-7" "-9999"
##   .. ..- attr(*, "names")= chr [1:4] "Missing" "Inconsistent data" "HaB short" "Consent withdrawn"
##  $ b650              : Factor w/ 5 levels "Consent withdrawn",..: NA NA 4 NA 3 3 3 3 3 3 ...
##  $ b663              : Factor w/ 8 levels "Consent withdrawn",..: NA NA 3 NA 4 4 4 3 4 3 ...
##  $ b665              : Factor w/ 8 levels "Consent withdrawn",..: NA NA 3 NA 4 4 4 3 4 3 ...
##  $ b667              : Factor w/ 8 levels "Consent withdrawn",..: NA NA 3 NA 4 4 4 3 4 3 ...
```

```
##  $ b862              : num  NA NA 10.9 NA 1.9 ...
##   ..- attr(*, "value.labels")= Named chr [1:2] "-1" "-9999"
##   .. ..- attr(*, "names")= chr [1:2] "Missing" "Consent withdrawn"
##  $ c645a             : Factor w/ 8 levels "Consent withdrawn",..: NA NA 7 4 4 4 6 5 4 8 ...
##  $ c755              : Factor w/ 9 levels "Consent withdrawn",..: NA NA 4 5 4 NA NA 5 7 5 ...
##  $ c765              : Factor w/ 9 levels "Consent withdrawn",..: NA NA 3 4 6 NA NA 6 6 3 ...
##  $ c800              : Factor w/ 12 levels "Consent withdrawn",..: NA NA NA 3 3 3 3 3 3 3 ...
##  $ c804              : Factor w/ 4 levels "Consent withdrawn",..: NA NA NA 3 3 3 3 3 3 3 ...
##  $ dw042             : num  NA NA 19.3 NA 21.3 ...
##   ..- attr(*, "value.labels")= Named chr [1:2] "-3" "-9999"
##   .. ..- attr(*, "names")= chr [1:2] "Missing" "Consent withdrawn"
##  $ bestgest          : num  NA NA 39 40 38 36 38 41 41 41 ...
##  - attr(*, "variable.labels")= Named chr [1:58] "Unique pregnancy identifier for Stine Marie Ulven(A
##   ..- attr(*, "names")= chr [1:58] "cidB3777" "qlet" "areases_quint_preg" "no2_preg" ...
##  - attr(*, "codepage")= int 1252
```

Several variables such as cholesterols are factors and should be integers.

```
alspac = alspac %>%
  mutate(across(c("chol_cord","trig_cord","HDL_cord","LDL_cord","CRP_cord",
                  "Trig_CIF31","Chol_CIF31","LDL_CIF31","HDL_CIF31",
                  "Trig_CIF43","Chol_CIF43","LDL_CIF43","HDL_CIF43",
                  "Chol_BBS","Trig_BBS","LDL_BBS","CHOL_F7","TRIG_F7","HDL_F7",
                  "LDL_F7","CHOL_F9","trig_f9","HDL_f9","LDL_f9","VLDL_f9","CRP_f9","IL6_f9"),as.numeri
str(alspac)
```

```
## 'data.frame':    15646 obs. of  58 variables:
##  $ cidB3777          : num  1 2 3 4 6 7 8 10 12 13 ...
##  $ qlet              : chr  "A" "A" "A" "A" ...
##  $ areases_quint_preg: num  NA NA 1 3 1 5 5 5 5 3 ...
##  $ no2_preg          : num  NA NA 22 26.9 29.6 26.5 27.9 27.4 33.3 27.9 ...
##  $ pm25_preg         : num  NA NA 14.5 13.5 13.3 13.4 13.4 13 14.5 13.8 ...
##  $ pm10_preg         : num  NA NA 27.8 29.8 NA ...
##  $ kz011b            : Factor w/ 4 levels "Consent withdrawn",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ kz021             : Factor w/ 4 levels "Consent withdrawn",..: 4 4 4 3 3 3 4 4 4 4 ...
##  $ kz030             : num  NA NA 3520 3570 3700 3340 2740 2900 3970 3600 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-10" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Not in core sample" "Triplet / quadruplet" "Consent withdrawn"
##  $ chol_cord         : num  NA NA NA 0.77 NA 1.46 NA 1.85 2.06 NA ...
##  $ trig_cord         : num  NA NA NA 0.41 NA 0.34 NA 0.7 0.63 NA ...
##  $ HDL_cord          : num  NA NA NA 0.23 NA 0.35 NA 0.38 0.74 NA ...
##  $ LDL_cord          : num  NA NA NA 0.354 NA ...
##  $ CRP_cord          : num  NA NA NA 0.06 NA 0 NA 0 0 NA ...
##  $ Trig_CIF31        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Chol_CIF31        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ LDL_CIF31         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ HDL_CIF31         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Trig_CIF43        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Chol_CIF43        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ LDL_CIF43         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ HDL_CIF43         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Chol_BBS          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Trig_BBS          : num  NA NA NA NA NA NA NA NA NA NA ...
```

```
##  $ HDL_BBS           : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing/Not assayed" "Trip/Quad" "Consent withdrawn"
##  $ LDL_BBS           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ CHOL_F7           : num  5.75 NA NA NA NA NA NA NA NA 4.37 ...
##  $ TRIG_F7           : num  1.83 NA NA NA NA NA NA NA NA 1.37 ...
##  $ HDL_F7            : num  1.26 NA NA NA NA NA NA NA NA 1.69 ...
##  $ LDL_F7            : num  3.61 NA NA NA NA NA NA NA NA 2.06 ...
##  $ CHOL_F9           : num  5.42 NA 4.09 NA 3.83 NA NA NA NA 3.98 ...
##  $ trig_f9           : num  0.74 NA 1.61 NA 0.37 NA NA NA NA 0.91 ...
##  $ HDL_f9            : num  1.47 NA 1.21 NA 1.67 NA NA NA NA 1.61 ...
##  $ LDL_f9            : num  3.61 NA 2.14 NA 1.99 ...
##  $ VLDL_f9           : num  0.338 NA 0.735 NA 0.169 ...
##  $ CRP_f9            : num  0.28 NA 0.11 NA 0.22 NA NA NA NA 0.58 ...
##  $ IL6_f9            : num  0.901 NA 1.139 NA 0.662 ...
##  $ age_31m           : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-2" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing" "Did not attend" "Consent withdrawn"
##  $ age_43m           : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-1" "-2" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Missing" "Did not attend" "Consent withdrawn"
##  $ age_7y            : num  7.65 9.15 NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr "-9999"
##   .. ..- attr(*, "names")= chr "Consent withdrawn"
##  $ age_8y            : num  8.42 10.03 9.28 NA 8.76 ...
##   ..- attr(*, "value.labels")= Named chr "-9999"
##   .. ..- attr(*, "names")= chr "Consent withdrawn"
##  $ age_9y            : num  9.67 10.15 10.3 NA 9.88 ...
##   ..- attr(*, "value.labels")= Named chr "-9999"
##   .. ..- attr(*, "names")= chr "Consent withdrawn"
##  $ mz010a            : Factor w/ 4 levels "Consent withdrawn",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ mz028b            : num  NA NA 33 23 29 31 18 25 25 29 ...
##   ..- attr(*, "value.labels")= Named chr [1:8] "44" "15" "-1" "-2" ...
##   .. ..- attr(*, "names")= chr [1:8] ">43" "< 16" "Ma's DOB NK" "Miscarried" ...
##  $ a525              : Factor w/ 9 levels "Consent withdrawn",..: NA NA 7 NA 6 7 3 3 7 7 ...
##  $ b032              : num  NA NA 1 NA 1 0 0 0 0 0 ...
##   ..- attr(*, "value.labels")= Named chr [1:4] "-1" "-2" "-7" "-9999"
##   .. ..- attr(*, "names")= chr [1:4] "Missing" "Inconsistent data" "HaB short" "Consent withdrawn"
##  $ b650              : Factor w/ 5 levels "Consent withdrawn",..: NA NA 4 NA 3 3 3 3 3 3 ...
##  $ b663              : Factor w/ 8 levels "Consent withdrawn",..: NA NA 3 NA 4 4 4 3 4 3 ...
##  $ b665              : Factor w/ 8 levels "Consent withdrawn",..: NA NA 3 NA 4 4 4 3 4 3 ...
##  $ b667              : Factor w/ 8 levels "Consent withdrawn",..: NA NA 3 NA 4 4 4 3 4 3 ...
##  $ b862              : num  NA NA 10.9 NA 1.9 ...
##   ..- attr(*, "value.labels")= Named chr [1:2] "-1" "-9999"
##   .. ..- attr(*, "names")= chr [1:2] "Missing" "Consent withdrawn"
##  $ c645a             : Factor w/ 8 levels "Consent withdrawn",..: NA NA 7 4 4 4 6 5 4 8 ...
##  $ c755              : Factor w/ 9 levels "Consent withdrawn",..: NA NA 4 5 4 NA NA 5 7 5 ...
##  $ c765              : Factor w/ 9 levels "Consent withdrawn",..: NA NA 3 4 6 NA NA 6 6 3 ...
##  $ c800              : Factor w/ 12 levels "Consent withdrawn",..: NA NA NA 3 3 3 3 3 3 3 ...
##  $ c804              : Factor w/ 4 levels "Consent withdrawn",..: NA NA NA 3 3 3 3 3 3 3 ...
##  $ dw042             : num  NA NA 19.3 NA 21.3 ...
##   ..- attr(*, "value.labels")= Named chr [1:2] "-3" "-9999"
##   .. ..- attr(*, "names")= chr [1:2] "Missing" "Consent withdrawn"
##  $ bestgest          : num  NA NA 39 40 38 36 38 41 41 41 ...
```

```
##  - attr(*, "variable.labels")= Named chr [1:58] "Unique pregnancy identifier for Stine Marie Ulven(A
##   ..- attr(*, "names")= chr [1:58] "cidB3777" "qlet" "areases_quint_preg" "no2_preg" ...
##  - attr(*, "codepage")= int 1252
```

Drop extra factors from variables that are categorical

Count the NA values per sample

```
na_count = apply(alspac,2,function(x){sum(is.na(x))})
na_count_o = sort(na_count,decreasing = TRUE)
na_count_o[1:10]
```

```
##  HDL_CIF31  LDL_CIF31 Trig_CIF31 Chol_CIF31  HDL_CIF43 Trig_CIF43 Chol_CIF43
##      15278      15274      15150      15150      15037      15033      15033
##  LDL_CIF43   Chol_BBS   Trig_BBS
##      15033      14778      14778
```

```
varibles_half_na = names(na_count_o)[na_count_o >= max(na_count_o)/2]
`%ni%` = Negate(`%in%`)
alspac_use = alspac[,colnames(alspac) %ni% varibles_half_na]
```

Using mice for missing data analysis among remaining variables

```
library(mice)
```

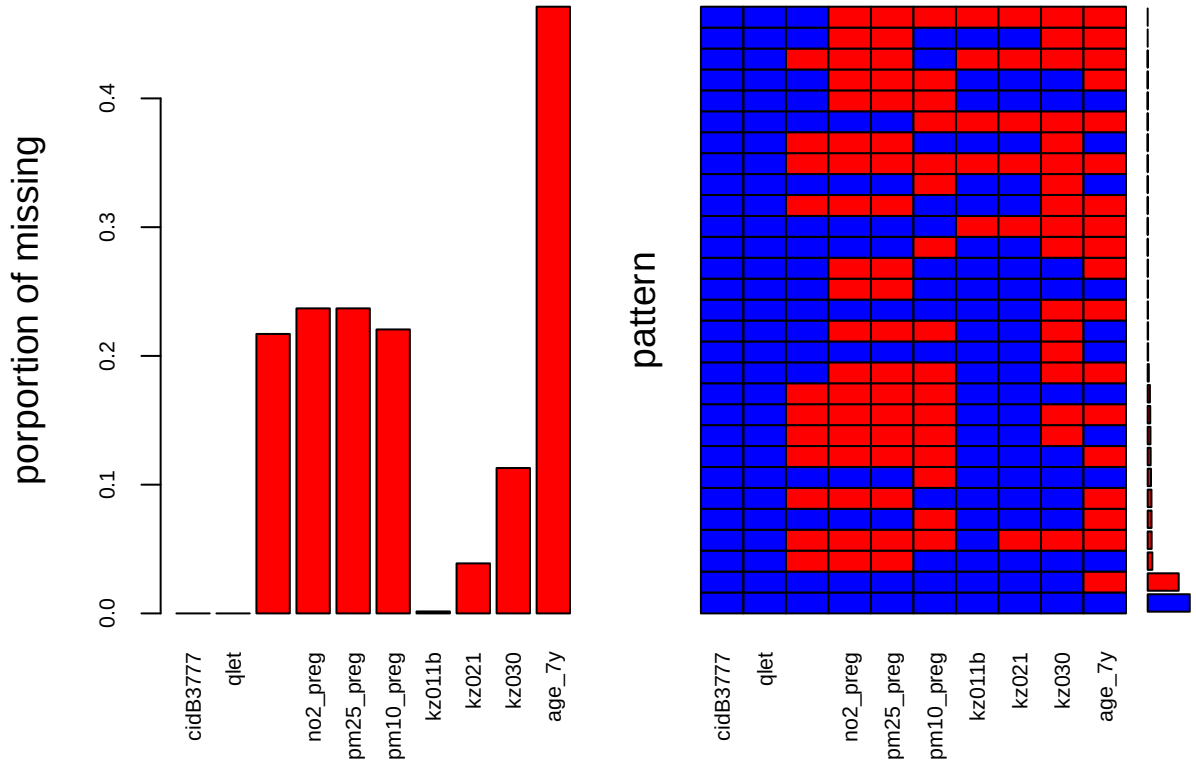
```
##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(VIM)
```

```
## Loading required package: colorspace

## Loading required package: grid

## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep
```

```
## red is missing and blue is not missing
missing_val_plot = aggr(alspac_use[,c(1:10)], col=c("blue","red"),
                    numbers=TRUE,sortVard=TRUE,
                      labels = names(alspac_use),cex.axis=.7,
                        gap=3,ylab=c("porportion of missing","pattern"))
```

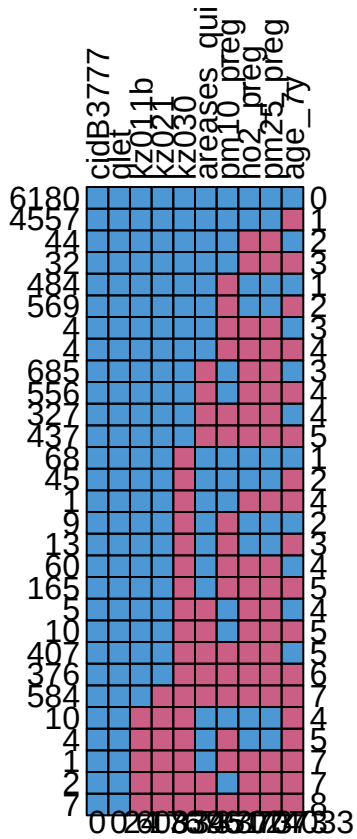
```
## Warning in plot.aggr(res, ...): not enough vertical space to display
## frequencies (too many combinations)
```



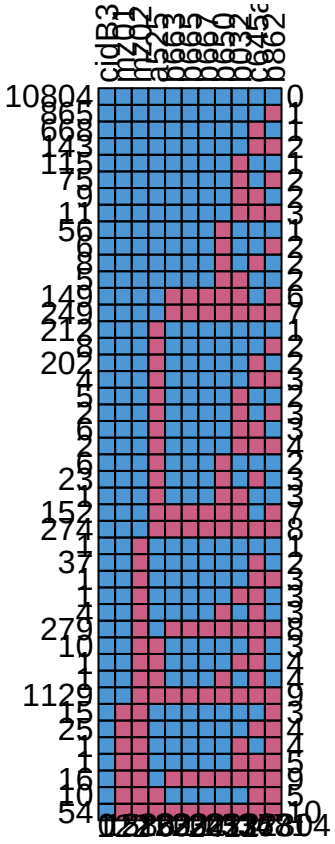

```
#md.pattern(alspac_use,rotate.names = TRUE)
```

Display missing-data patterns

```
tab.pattern=md.pattern(alspac_use[,c(1:10)],rotate.names = TRUE) ## difficul tot understand
```



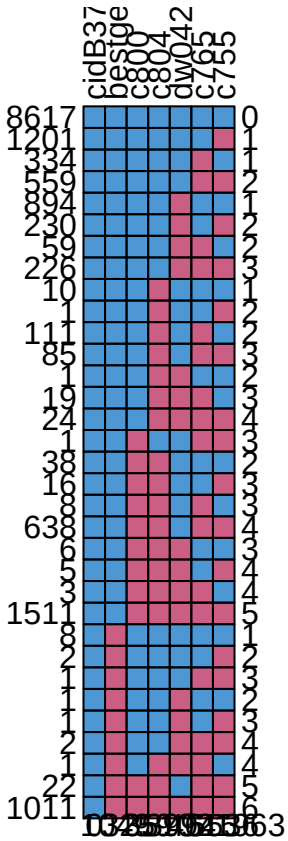

```
md.pattern(alspac_use[,c(1,11:20)],rotate.names = TRUE)
```



```
##       cidB3777 mz010a mz028b a525 b663 b665 b667 b650 b032 c645a b862
## 10804        1      1      1    1    1    1    1    1    1     1    1     0
## 865          1      1      1    1    1    1    1    1    1     1    0     1
## 668          1      1      1    1    1    1    1    1    1     0    1     1
## 143          1      1      1    1    1    1    1    1    1     0    0     2
## 115          1      1      1    1    1    1    1    1    0     1    1     1
## 75           1      1      1    1    1    1    1    1    0     1    0     2
## 9            1      1      1    1    1    1    1    1    0     0    1     2
## 11           1      1      1    1    1    1    1    1    0     0    0     3
## 56           1      1      1    1    1    1    1    0    1     1    1     1
## 6            1      1      1    1    1    1    1    0    1     1    0     2
## 8            1      1      1    1    1    1    1    0    1     0    1     2
## 5            1      1      1    1    1    1    1    0    0     1    1     2
## 149          1      1      1    1    0    0    0    0    0     1    0     6
## 249          1      1      1    1    0    0    0    0    0     0    0     7
## 212          1      1      1    0    1    1    1    1    1     1    1     1
## 8            1      1      1    0    1    1    1    1    1     1    0     2
## 202          1      1      1    0    1    1    1    1    1     0    1     2
## 4            1      1      1    0    1    1    1    1    1     0    0     3
## 5            1      1      1    0    1    1    1    1    0     1    1     2
## 2            1      1      1    0    1    1    1    1    0     1    0     3
## 6            1      1      1    0    1    1    1    1    0     0    1     3
## 2            1      1      1    0    1    1    1    1    0     0    0     4
## 6            1      1      1    0    1    1    1    0    1     1    1     2
## 23           1      1      1    0    1    1    1    0    1     0    1     3
## 1            1      1      1    0    1    1    1    0    0     1    1     3
```

```
## 152          1       1       1     0     0     0     0     0     0      1     0      7
## 274          1       1       1     0     0     0     0     0     0      0     0      8
## 1            1       1       0     1     1     1     1     1     1      1     1      1
## 37           1       1       0     1     1     1     1     1     1      0     1      2
## 1            1       1       0     1     1     1     1     1     1      0     0      3
## 1            1       1       0     1     1     1     1     1     0      0     1      3
## 4            1       1       0     1     1     1     1     0     1      0     1      3
## 279          1       1       0     1     0     0     0     0     0      0     0      8
## 10           1       1       0     0     1     1     1     1     1      0     1      3
## 1            1       1       0     0     1     1     1     1     0      0     1      4
## 1            1       1       0     0     1     1     1     0     1      0     1      4
## 1129         1       1       0     0     0     0     0     0     0      0     0      9
## 15           1       0       0     1     1     1     1     1     1      1     0      3
## 25           1       0       0     1     1     1     1     1     1      0     0      4
## 1            1       0       0     1     1     1     1     1     0      1     0      4
## 1            1       0       0     1     1     1     1     1     0      0     0      5
## 16           1       0       0     1     0     0     0     0     0      0     0      9
## 10           1       0       0     0     1     1     1     1     1      0     0      5
## 54           1       0       0     0     0     0     0     0     0      0     0     10
##              0     122    1586  2102  2302  2302  2302  2412  2537   3168  3471  22304
```

```
md.pattern(alspac_use[,c(1,21:26)],rotate.names = TRUE)
```




```
##      cidB3777 bestgest c800 c804 dw042 c765 c755
## 8617        1        1    1    1     1    1    1     0
```

```
## 1201        1        1    1    1     1    1    0     1
## 334         1        1    1    1     1    0    1     1
## 559         1        1    1    1     1    0    0     2
## 894         1        1    1    1     0    1    1     1
## 230         1        1    1    1     0    1    0     2
## 59          1        1    1    1     0    0    1     2
## 226         1        1    1    1     0    0    0     3
## 10          1        1    1    0     1    1    1     1
## 1           1        1    1    0     1    1    0     2
## 111         1        1    1    0     1    0    1     2
## 85          1        1    1    0     1    0    0     3
## 1           1        1    1    0     0    1    1     2
## 19          1        1    1    0     0    0    1     3
## 24          1        1    1    0     0    0    0     4
## 1           1        1    0    1     1    0    0     3
## 38          1        1    0    0     1    1    1     2
## 16          1        1    0    0     1    1    0     3
## 8           1        1    0    0     1    0    1     3
## 638         1        1    0    0     1    0    0     4
## 6           1        1    0    0     0    1    1     3
## 5           1        1    0    0     0    1    0     4
## 3           1        1    0    0     0    0    1     4
## 1511        1        1    0    0     0    0    0     5
## 8           1        0    1    1     1    1    1     1
## 2           1        0    1    1     1    1    0     2
## 1           1        0    1    1     1    0    0     3
## 1           1        0    1    1     0    1    1     2
## 1           1        0    1    1     0    1    0     3
## 2           1        0    1    1     0    0    0     4
## 1           1        0    1    0     0    0    1     4
## 22          1        0    0    0     1    0    0     5
## 1011        1        0    0    0     0    0    0     6
##             0     1049 3259 3510  3994 4615 5536 21963
```

Doing missing data pattern for the most interesting variables

```
cols_use = c("cidB3777","no2_preg","pm25_preg","pm10_preg","kz021","c800","c804","c645a","b650")
pdf("Missing_pattern_table.pdf",width = 10,height = 10)
md.pattern(alspac_use[,cols_use],rotate.names = TRUE)
```

```
##      cidB3777 kz021 b650 c645a c800 pm10_preg c804 no2_preg pm25_preg
## 9168        1     1    1     1    1         1    1        1         1     0
## 1037        1     1    1     1    1         1    1        0         0     2
## 168         1     1    1     1    1         1    0        1         1     1
## 25          1     1    1     1    1         1    0        0         0     3
## 861         1     1    1     1    1         0    1        1         1     1
## 644         1     1    1     1    1         0    1        0         0     3
## 28          1     1    1     1    1         0    0        1         1     2
## 12          1     1    1     1    1         0    0        0         0     4
## 1           1     1    1     1    0         1    1        1         1     1
## 108         1     1    1     1    0         1    0        1         1     2
## 12          1     1    1     1    0         1    0        0         0     4
## 11          1     1    1     1    0         0    0        1         1     3
```

```
## 11          1     1    1     1    0         0    0        0         0     5
## 43          1     1    1     0    1         1    1        1         1     1
## 13          1     1    1     0    1         1    1        0         0     3
## 3           1     1    1     0    1         1    0        1         1     2
## 1           1     1    1     0    1         1    0        0         0     4
## 5           1     1    1     0    1         0    1        1         1     2
## 2           1     1    1     0    1         0    1        0         0     4
## 700         1     1    1     0    0         1    0        1         1     3
## 117         1     1    1     0    0         1    0        0         0     5
## 77          1     1    1     0    0         0    0        1         1     4
## 111         1     1    1     0    0         0    0        0         0     6
## 243         1     1    0     1    1         1    1        1         1     1
## 51          1     1    0     1    1         1    1        0         0     3
## 7           1     1    0     1    1         1    0        1         1     2
## 4           1     1    0     1    1         1    0        0         0     4
## 31          1     1    0     1    1         0    1        1         1     2
## 18          1     1    0     1    1         0    1        0         0     4
## 1           1     1    0     1    1         0    0        1         1     3
## 12          1     1    0     1    0         1    0        1         1     3
## 5           1     1    0     1    0         1    0        0         0     5
## 3           1     1    0     1    0         0    0        1         1     4
## 3           1     1    0     0    1         1    1        1         1     2
## 2           1     1    0     0    1         1    0        1         1     3
## 1           1     1    0     0    1         0    1        1         1     3
## 392         1     1    0     0    0         1    0        1         1     4
## 68          1     1    0     0    0         1    0        0         0     6
## 57          1     1    0     0    0         0    0        1         1     5
## 982         1     1    0     0    0         0    0        0         0     7
## 15          1     0    1     1    1         0    1        0         0     4
## 1           1     0    1     1    1         0    0        0         0     5
## 1           1     0    1     1    0         0    0        0         0     6
## 59          1     0    1     0    0         0    0        0         0     7
## 10          1     0    0     0    0         1    0        1         1     5
## 2           1     0    0     0    0         1    0        0         0     7
## 4           1     0    0     0    0         0    0        1         1     6
## 516         1     0    0     0    0         0    0        0         0     8
##             0   608 2412  3168 3259      3451 3510     3707      3707 23822
```

```
dev.off()
```

```
## pdf
##   2
```

Both no2_preg and pm25_preg have the most missing values

```
missing_b650<-is.na(alspac_use$b650)#Missing indicator for ever smoked

missing_c645a<-is.na(alspac_use$c645a)#Missing indicator for mom education

missing_c800<-is.na(alspac_use$c800)#Missing indicator for ethnic group

missing_c804<-is.na(alspac_use$c804)#Missing indicator for child ethnicity

missing_areases<-is.na(alspac_use$areases_quint_preg)#Missing indicator for areaSES
```

Using the missing data indicator variable created, let us compare the mean PM2.5 exposure of those with missing values on Length against those whose Length was observed.

```
tapply(X=alspac_use$pm25_preg,INDEX=missing_b650,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 13.33947 13.35979
```

```
tapply(X=alspac_use$pm25_preg,INDEX=missing_c645a,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 13.33805 13.36315
```

```
tapply(X=alspac_use$pm25_preg,INDEX=missing_c800,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 13.33840 13.35898
```

```
tapply(X=alspac_use$pm25_preg,INDEX=missing_c804,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 13.33690 13.36614
```

```
tapply(X=alspac_use$pm25_preg,INDEX=missing_areases,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 13.34077      NaN
```

Above table suggests no notable difference pm25_preg for those categorical variables.

```
t.test(alspac_use$pm25_preg ~ missing_b650)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use$pm25_preg by missing_b650
## t = -0.69963, df = 897.66, p-value = 0.4843
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.07732555  0.03668361
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            13.33947            13.35979
```

```
t.test(alspac_use$pm25_preg ~ missing_c645a)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use$pm25_preg by missing_c645a
## t = -1.0315, df = 1650.2, p-value = 0.3025
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.07282337  0.02262665
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            13.33805            13.36315
```

```
t.test(alspac_use$pm25_preg ~ missing_c800)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use$pm25_preg by missing_c800
## t = -0.86792, df = 1777, p-value = 0.3856
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.06707896  0.02592335
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            13.33840            13.35898
```

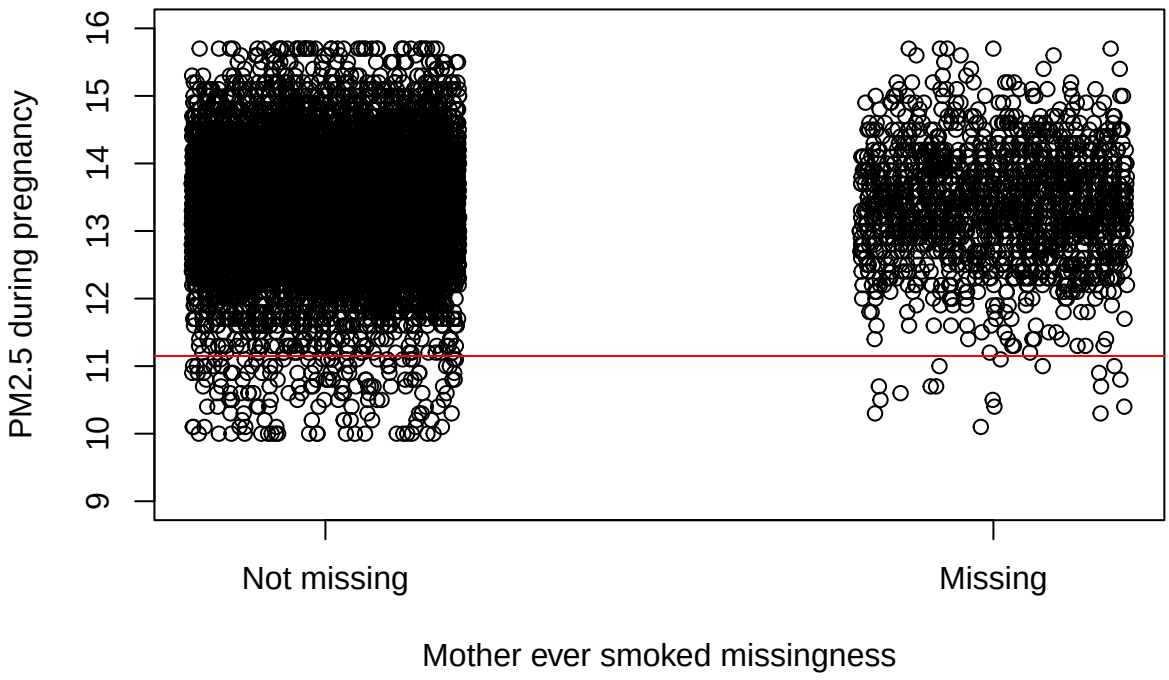
```
t.test(alspac_use$pm25_preg ~ missing_c804)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use$pm25_preg by missing_c804
## t = -1.3094, df = 2131.1, p-value = 0.1905
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.07304217  0.01455466
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            13.33690            13.36614
```

```
#t.test(alspac_use$pm25_preg ~ missing_areases)
```

There is no significant difference in PM2.5 exposure for those with or without data on some potential covariates. Making a jitter plot for PM2.5 by various groups

```
plot(jitter(as.numeric(missing_c804)),alspac_use$pm25_preg,ylim=c(9,16),
            xaxt="n",
            xlab="Mother ever smoked missingness",
            ylab="PM2.5 during pregnancy")
axis(1, at=c(0,1), labels=c("Not missing","Missing"))

abline(h=quantile(alspac_use$pm25_preg,1/4,na.rm=T)-1.5*IQR(alspac_use$pm25_preg,na.rm=T),col="red")
```



Now doing the same for PM10 First checking number of missing Pm10 values by confounders

```
tapply(X=alspac_use$pm10_preg,INDEX=missing_b650,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 31.14124 31.47441
```

```
tapply(X=alspac_use$pm10_preg,INDEX=missing_c645a,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 31.13799 31.36386
```

```
tapply(X=alspac_use$pm10_preg,INDEX=missing_c800,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 31.13199 31.39760
```

```
tapply(X=alspac_use$pm10_preg,INDEX=missing_c804,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 31.13187 31.36441
```

```
tapply(X=alspac_use$pm10_preg,INDEX=missing_areases,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 31.16037 31.18655
```

T test for PM10

```
t.test(alspac_use$pm10_preg ~ missing_b650)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use$pm10_preg by missing_b650
## t = -4.6156, df = 922.84, p-value = 4.475e-06
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.4748326 -0.1915051
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            31.14124            31.47441
```

```
t.test(alspac_use$pm10_preg ~ missing_c645a)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use$pm10_preg by missing_c645a
## t = -3.8678, df = 1714.7, p-value = 0.0001139
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.3404006 -0.1113307
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            31.13799            31.36386
```

```
t.test(alspac_use$pm10_preg ~ missing_c800)
```

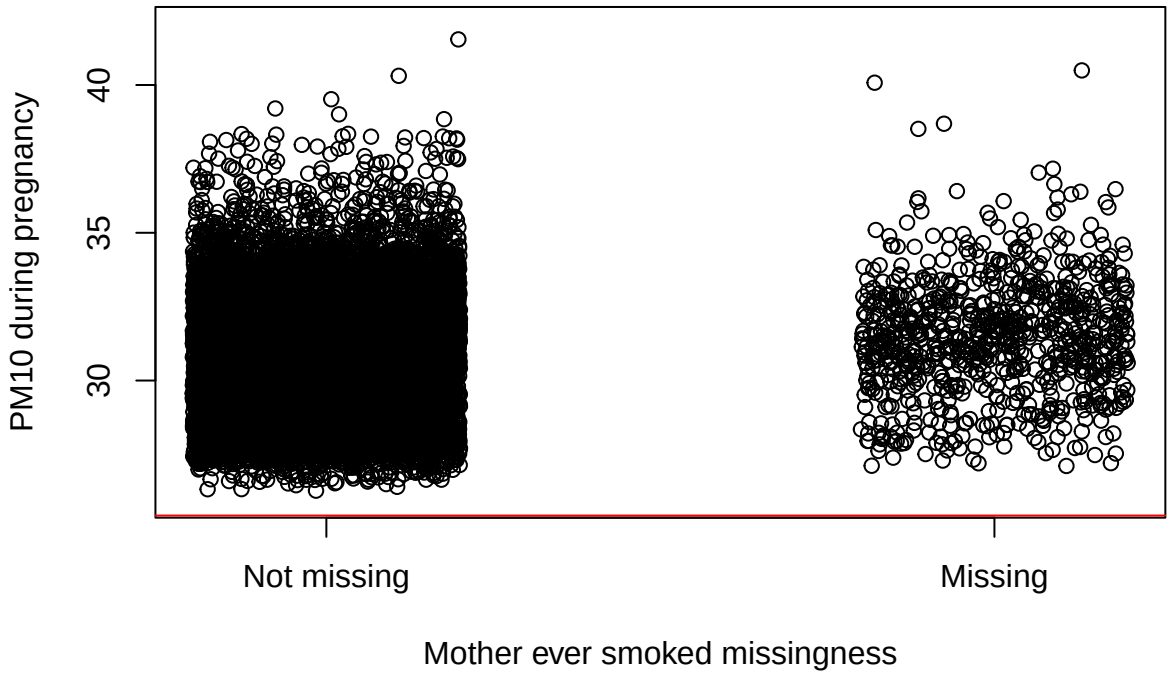
```
##
##  Welch Two Sample t-test
##
## data:  alspac_use$pm10_preg by missing_c800
## t = -4.7115, df = 1845.2, p-value = 2.643e-06
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.3761749 -0.1550434
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            31.13199            31.39760
```

```r
t.test(alspac_use$pm10_preg ~ missing_c804)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use$pm10_preg by missing_c804
## t = -4.3748, df = 2202.5, p-value = 1.272e-05
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.3367764 -0.1282999
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            31.13187            31.36441
```

```r
#t.test(alspac_use$pm25_preg ~ missing_areases)
```

```r
plot(jitter(as.numeric(missing_b650)),alspac_use$pm10_preg,ylim=c(26,42),
            xaxt="n",
            xlab="Mother ever smoked missingness",
            ylab="PM10 during pregnancy")
axis(1, at=c(0,1), labels=c("Not missing","Missing"))

abline(h=quantile(alspac_use$pm10_preg,1/4,na.rm=T)-1.5*IQR(alspac_use$pm10_preg,na.rm=T),col="red")
```



Now, I will drop the extra levels of all categorical variables that do not have values corresponding to those levels

```
alspac_use_no_extra_factor <- alspac_use %>%
  mutate(across(where(is.factor), fct_drop))
```

Now chekcing if extra factors dropped

```
str(alspac_use)
```

```
## 'data.frame':    15646 obs. of  26 variables:
##  $ cidB3777          : num  1 2 3 4 6 7 8 10 12 13 ...
##  $ qlet              : chr  "A" "A" "A" "A" ...
##  $ areases_quint_preg: num  NA NA 1 3 1 5 5 5 5 3 ...
##  $ no2_preg          : num  NA NA 22 26.9 29.6 26.5 27.9 27.4 33.3 27.9 ...
##  $ pm25_preg         : num  NA NA 14.5 13.5 13.3 13.4 13.4 13 14.5 13.8 ...
##  $ pm10_preg         : num  NA NA 27.8 29.8 NA ...
##  $ kz011b            : Factor w/ 4 levels "Consent withdrawn",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ kz021             : Factor w/ 4 levels "Consent withdrawn",..: 4 4 4 3 3 3 4 4 4 4 ...
##  $ kz030             : num  NA NA 3520 3570 3700 3340 2740 2900 3970 3600 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-10" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Not in core sample" "Triplet / quadruplet" "Consent withdrawn"
##  $ age_7y            : num  7.65 9.15 NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr "-9999"
##   .. ..- attr(*, "names")= chr "Consent withdrawn"
##  $ mz010a            : Factor w/ 4 levels "Consent withdrawn",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ mz028b            : num  NA NA 33 23 29 31 18 25 25 29 ...
##   ..- attr(*, "value.labels")= Named chr [1:8] "44" "15" "-1" "-2" ...
##   .. ..- attr(*, "names")= chr [1:8] ">43" "< 16" "Ma's DOB NK" "Miscarried" ...
##  $ a525              : Factor w/ 9 levels "Consent withdrawn",..: NA NA 7 NA 6 7 3 3 7 7 ...
##  $ b032              : num  NA NA 1 NA 1 0 0 0 0 0 ...
##   ..- attr(*, "value.labels")= Named chr [1:4] "-1" "-2" "-7" "-9999"
##   .. ..- attr(*, "names")= chr [1:4] "Missing" "Inconsistent data" "HaB short" "Consent withdrawn"
##  $ b650              : Factor w/ 5 levels "Consent withdrawn",..: NA NA 4 NA 3 3 3 3 3 3 ...
##  $ b663              : Factor w/ 8 levels "Consent withdrawn",..: NA NA 3 NA 4 4 4 3 4 3 ...
##  $ b665              : Factor w/ 8 levels "Consent withdrawn",..: NA NA 3 NA 4 4 4 3 4 3 ...
##  $ b667              : Factor w/ 8 levels "Consent withdrawn",..: NA NA 3 NA 4 4 4 3 4 3 ...
##  $ b862              : num  NA NA 10.9 NA 1.9 ...
##   ..- attr(*, "value.labels")= Named chr [1:2] "-1" "-9999"
##   .. ..- attr(*, "names")= chr [1:2] "Missing" "Consent withdrawn"
##  $ c645a             : Factor w/ 8 levels "Consent withdrawn",..: NA NA 7 4 4 4 6 5 4 8 ...
##  $ c755              : Factor w/ 9 levels "Consent withdrawn",..: NA NA 4 5 4 NA NA 5 7 5 ...
##  $ c765              : Factor w/ 9 levels "Consent withdrawn",..: NA NA 3 4 6 NA NA 6 6 3 ...
##  $ c800              : Factor w/ 12 levels "Consent withdrawn",..: NA NA NA 3 3 3 3 3 3 3 ...
##  $ c804              : Factor w/ 4 levels "Consent withdrawn",..: NA NA NA 3 3 3 3 3 3 3 ...
##  $ dw042             : num  NA NA 19.3 NA 21.3 ...
##   ..- attr(*, "value.labels")= Named chr [1:2] "-3" "-9999"
##   .. ..- attr(*, "names")= chr [1:2] "Missing" "Consent withdrawn"
##  $ bestgest          : num  NA NA 39 40 38 36 38 41 41 41 ...
```

```
str(alspac_use_no_extra_factor)
```

```
## 'data.frame':    15646 obs. of  26 variables:
```

```
##  $ cidB3777          : num  1 2 3 4 6 7 8 10 12 13 ...
##  $ qlet              : chr  "A" "A" "A" "A" ...
##  $ areases_quint_preg: num  NA NA 1 3 1 5 5 5 5 3 ...
##  $ no2_preg          : num  NA NA 22 26.9 29.6 26.5 27.9 27.4 33.3 27.9 ...
##  $ pm25_preg         : num  NA NA 14.5 13.5 13.3 13.4 13.4 13 14.5 13.8 ...
##  $ pm10_preg         : num  NA NA 27.8 29.8 NA ...
##  $ kz011b            : Factor w/ 2 levels "Yes","No": 1 1 1 1 1 1 1 1 1 1 ...
##  $ kz021             : Factor w/ 2 levels "Male","Female": 2 2 2 1 1 1 2 2 2 2 ...
##  $ kz030             : num  NA NA 3520 3570 3700 3340 2740 2900 3970 3600 ...
##   ..- attr(*, "value.labels")= Named chr [1:3] "-10" "-11" "-9999"
##   .. ..- attr(*, "names")= chr [1:3] "Not in core sample" "Triplet / quadruplet" "Consent withdrawn"
##  $ age_7y            : num  7.65 9.15 NA NA NA ...
##   ..- attr(*, "value.labels")= Named chr "-9999"
##   .. ..- attr(*, "names")= chr "Consent withdrawn"
##  $ mz010a            : Factor w/ 2 levels "Singleton","Multiple": 1 1 1 1 1 1 1 1 1 1 ...
##  $ mz028b            : num  NA NA 33 23 29 31 18 25 25 29 ...
##   ..- attr(*, "value.labels")= Named chr [1:8] "44" "15" "-1" "-2" ...
##   .. ..- attr(*, "names")= chr [1:8] ">43" "< 16" "Ma's DOB NK" "Miscarried" ...
##  $ a525              : Factor w/ 6 levels "Never married",..: NA NA 5 NA 4 5 1 1 5 5 ...
##  $ b032              : num  NA NA 1 NA 1 0 0 0 0 0 ...
##   ..- attr(*, "value.labels")= Named chr [1:4] "-1" "-2" "-7" "-9999"
##   .. ..- attr(*, "names")= chr [1:4] "Missing" "Inconsistent data" "HaB short" "Consent withdrawn"
##  $ b650              : Factor w/ 2 levels "Y","N": NA NA 2 NA 1 1 1 1 1 1 ...
##  $ b663              : Factor w/ 5 levels "N","Y CIGS","Y cigars",..: NA NA 1 NA 2 2 2 1 2 1 ...
##  $ b665              : Factor w/ 4 levels "N","Y CIGS","Y cigars",..: NA NA 1 NA 2 2 2 1 2 1 ...
##  $ b667              : Factor w/ 4 levels "N","Y CIGS","Y cigars",..: NA NA 1 NA 2 2 2 1 2 1 ...
##  $ b862              : num  NA NA 10.9 NA 1.9 ...
##   ..- attr(*, "value.labels")= Named chr [1:2] "-1" "-9999"
##   .. ..- attr(*, "names")= chr [1:2] "Missing" "Consent withdrawn"
##  $ c645a             : Factor w/ 5 levels "CSE","Vocational",..: NA NA 4 1 1 1 3 2 1 5 ...
##  $ c755              : Factor w/ 7 levels "I","II","III (non-manual)",..: NA NA 2 3 2 NA NA 3 5 3 ..
##  $ c765              : Factor w/ 7 levels "I","II","III (non-manual)",..: NA NA 1 2 4 NA NA 4 4 1 ..
##  $ c800              : Factor w/ 9 levels "White","Black Caribbean",..: NA NA NA 1 1 1 1 1 1 1 ...
##  $ c804              : Factor w/ 2 levels "White","Non-white": NA NA NA 1 1 1 1 1 1 1 ...
##  $ dw042             : num  NA NA 19.3 NA 21.3 ...
##   ..- attr(*, "value.labels")= Named chr [1:2] "-3" "-9999"
##   .. ..- attr(*, "names")= chr [1:2] "Missing" "Consent withdrawn"
##  $ bestgest          : num  NA NA 39 40 38 36 38 41 41 41 ...
```

Using the missing data indicator variable created, let us compare the mean NO2 exposure of those with missing values on Length against those whose Length was observed.

```
tapply(X=alspac_use_no_extra_factor$no2_preg,INDEX=missing_b650,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 26.94640 27.58081
```

```
tapply(X=alspac_use_no_extra_factor$no2_preg,INDEX=missing_c645a,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 26.93880 27.38342
```

```
tapply(X=alspac_use_no_extra_factor$no2_preg,INDEX=missing_c800,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 26.93377 27.39687
```

```
tapply(X=alspac_use_no_extra_factor$no2_preg,INDEX=missing_c804,FUN=mean,na.rm=T)
```

```
##    FALSE     TRUE
## 26.92754 27.37675
```

```
#tapply(X=alspac_use$no2_preg,INDEX=missing_areases,FUN=mean,na.rm=T)
```

T-test for NO2.

```
t.test(alspac_use_no_extra_factor$no2_preg ~ missing_b650)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use_no_extra_factor$no2_preg by missing_b650
## t = -4.4196, df = 890.78, p-value = 1.111e-05
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.9161380 -0.3526856
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            26.94640            27.58081
```

```
t.test(alspac_use_no_extra_factor$no2_preg ~ missing_c645a)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use_no_extra_factor$no2_preg by missing_c645a
## t = -3.7479, df = 1642.1, p-value = 0.0001845
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.6773115 -0.2119369
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            26.93880            27.38342
```

```
t.test(alspac_use_no_extra_factor$no2_preg ~ missing_c800)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use_no_extra_factor$no2_preg by missing_c800
## t = -4.0339, df = 1775, p-value = 5.72e-05
```

```
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.6882748 -0.2379394
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            26.93377            27.39687
```

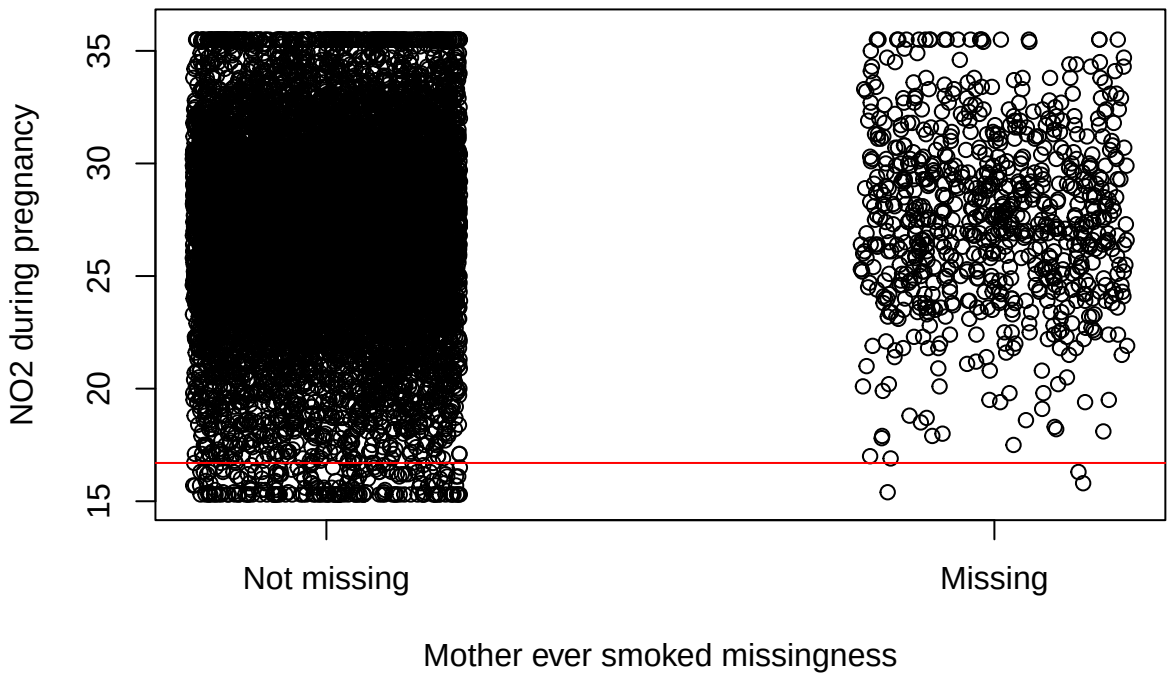
```
t.test(alspac_use_no_extra_factor$no2_preg ~ missing_c804)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use_no_extra_factor$no2_preg by missing_c804
## t = -4.1539, df = 2128.2, p-value = 3.397e-05
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.6612884 -0.2371384
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            26.92754            27.37675
```

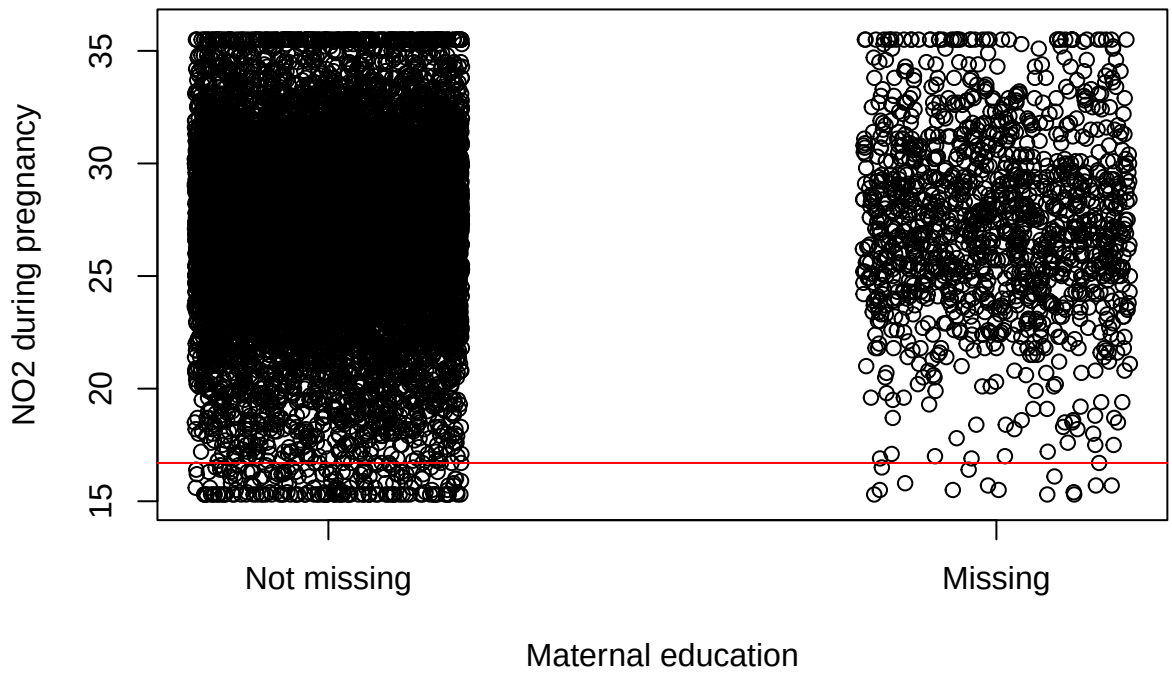
```
#t.test(alspac_use$pm25_preg ~ missing_areases)
```

The above data suggests that NO2 has different distribution by several of the potential covariates.

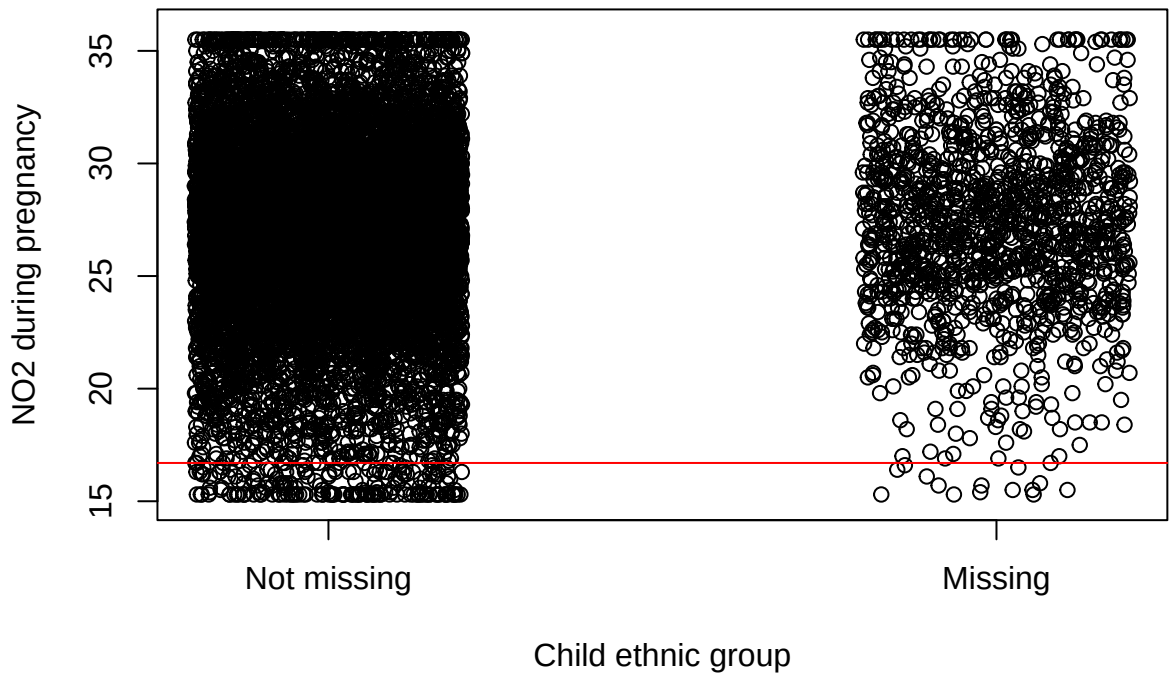
```
plot(jitter(as.numeric(missing_b650)),alspac_use$no2_preg,ylim=c(15,36),
            xaxt="n",
            xlab="Mother ever smoked missingness",
            ylab="NO2 during pregnancy")
axis(1, at=c(0,1), labels=c("Not missing","Missing"))

abline(h=quantile(alspac_use$no2_preg,1/4,na.rm=T)-1.5*IQR(alspac_use$no2_preg,na.rm=T),col="red")
```



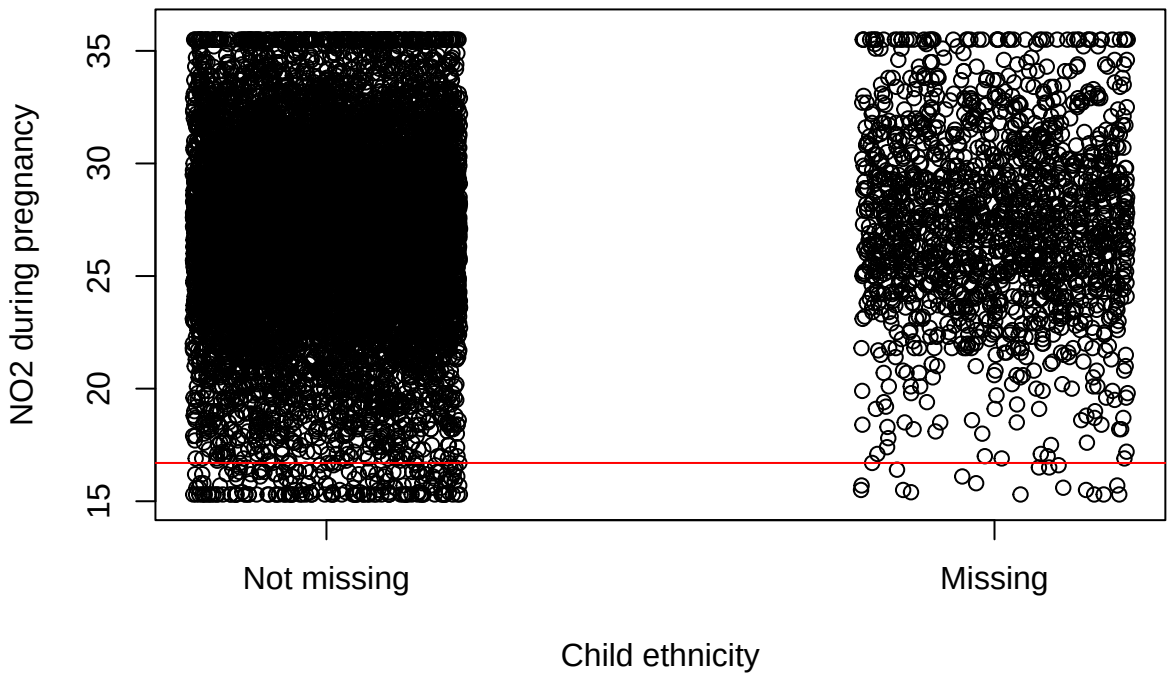

```
plot(jitter(as.numeric(missing_c645a)),alspac_use$no2_preg,ylim=c(15,36),
             xaxt="n",
             xlab="Maternal education",
             ylab="NO2 during pregnancy")
axis(1, at=c(0,1), labels=c("Not missing","Missing"))

abline(h=quantile(alspac_use$no2_preg,1/4,na.rm=T)-1.5*IQR(alspac_use$no2_preg,na.rm=T),col="red")
```



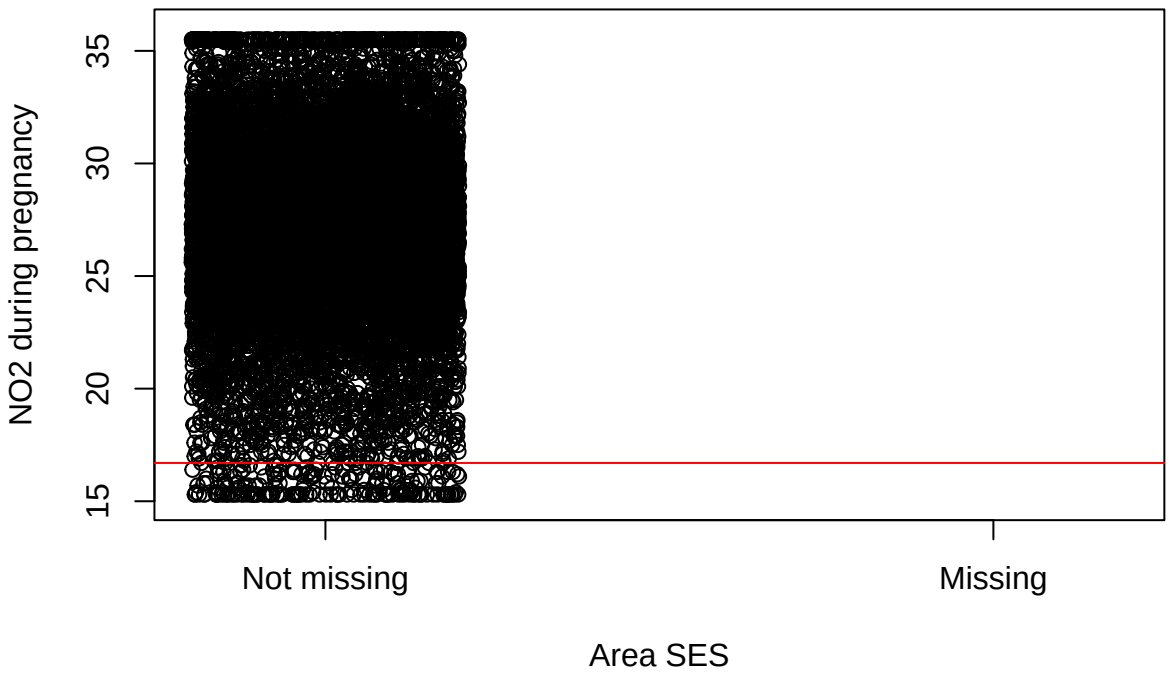

```
plot(jitter(as.numeric(missing_c800)),alspac_use$no2_preg,ylim=c(15,36),
             xaxt="n",
             xlab="Child ethnic group",
             ylab="NO2 during pregnancy")
axis(1, at=c(0,1), labels=c("Not missing","Missing"))

abline(h=quantile(alspac_use$no2_preg,1/4,na.rm=T)-1.5*IQR(alspac_use$no2_preg,na.rm=T),col="red")
```




```
plot(jitter(as.numeric(missing_c804)),alspac_use$no2_preg,ylim=c(15,36),
             xaxt="n",
             xlab="Child ethnicity",
             ylab="NO2 during pregnancy")
axis(1, at=c(0,1), labels=c("Not missing","Missing"))

abline(h=quantile(alspac_use$no2_preg,1/4,na.rm=T)-1.5*IQR(alspac_use$no2_preg,na.rm=T),col="red")
```




```
plot(jitter(as.numeric(missing_areases)),alspac_use$no2_preg,ylim=c(15,36),
             xaxt="n",
             xlab="Area SES",
             ylab="NO2 during pregnancy")
axis(1, at=c(0,1), labels=c("Not missing","Missing"))

abline(h=quantile(alspac_use$no2_preg,1/4,na.rm=T)-1.5*IQR(alspac_use$no2_preg,na.rm=T),col="red")
```



Checking if the exposures are different by gender

```
t.test(alspac_use_no_extra_factor$no2_preg ~ alspac_use_no_extra_factor$kz021)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use_no_extra_factor$no2_preg by alspac_use_no_extra_factor$kz021
## t = 0.70609, df = 11880, p-value = 0.4801
## alternative hypothesis: true difference in means between group Male and group Female is not equal to
## 95 percent confidence interval:
##  -0.0941551  0.2001807
## sample estimates:
##   mean in group Male mean in group Female
##             27.01492             26.96191
```

```
t.test(alspac_use_no_extra_factor$pm25_preg ~ alspac_use_no_extra_factor$kz021)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use_no_extra_factor$pm25_preg by alspac_use_no_extra_factor$kz021
## t = 1.0134, df = 11852, p-value = 0.3109
## alternative hypothesis: true difference in means between group Male and group Female is not equal to
## 95 percent confidence interval:
```

```
##  -0.01470921  0.04619499
## sample estimates:
##   mean in group Male mean in group Female
##             13.34865             13.33290
```

```
t.test(alspac_use_no_extra_factor$pm10_preg ~ alspac_use_no_extra_factor$kz021)
```

```
##
##  Welch Two Sample t-test
##
## data:  alspac_use_no_extra_factor$pm10_preg by alspac_use_no_extra_factor$kz021
## t = 0.30433, df = 12153, p-value = 0.7609
## alternative hypothesis: true difference in means between group Male and group Female is not equal to
## 95 percent confidence interval:
##  -0.06118993  0.08368208
## sample estimates:
##   mean in group Male mean in group Female
##             31.16920             31.15795
```

While distribution of pollutant exposure does not differ by gender, gender differences lead to differences in expression of genes and hence adjusting for gender is necessary. This analysis is followed by differential expression analysis using the bryois.csv microarray matrix from ALSPAC.