Case Study 1 -

Alessandro Zito

January 16, 2020

R setup

```
suppressMessages(library(tidyverse))
library(corrplot)

## corrplot 0.84 loaded

library(RColorBrewer)
ggplot2::theme_set(ggplot2::theme_bw())
```

Exercise 1 Introduction

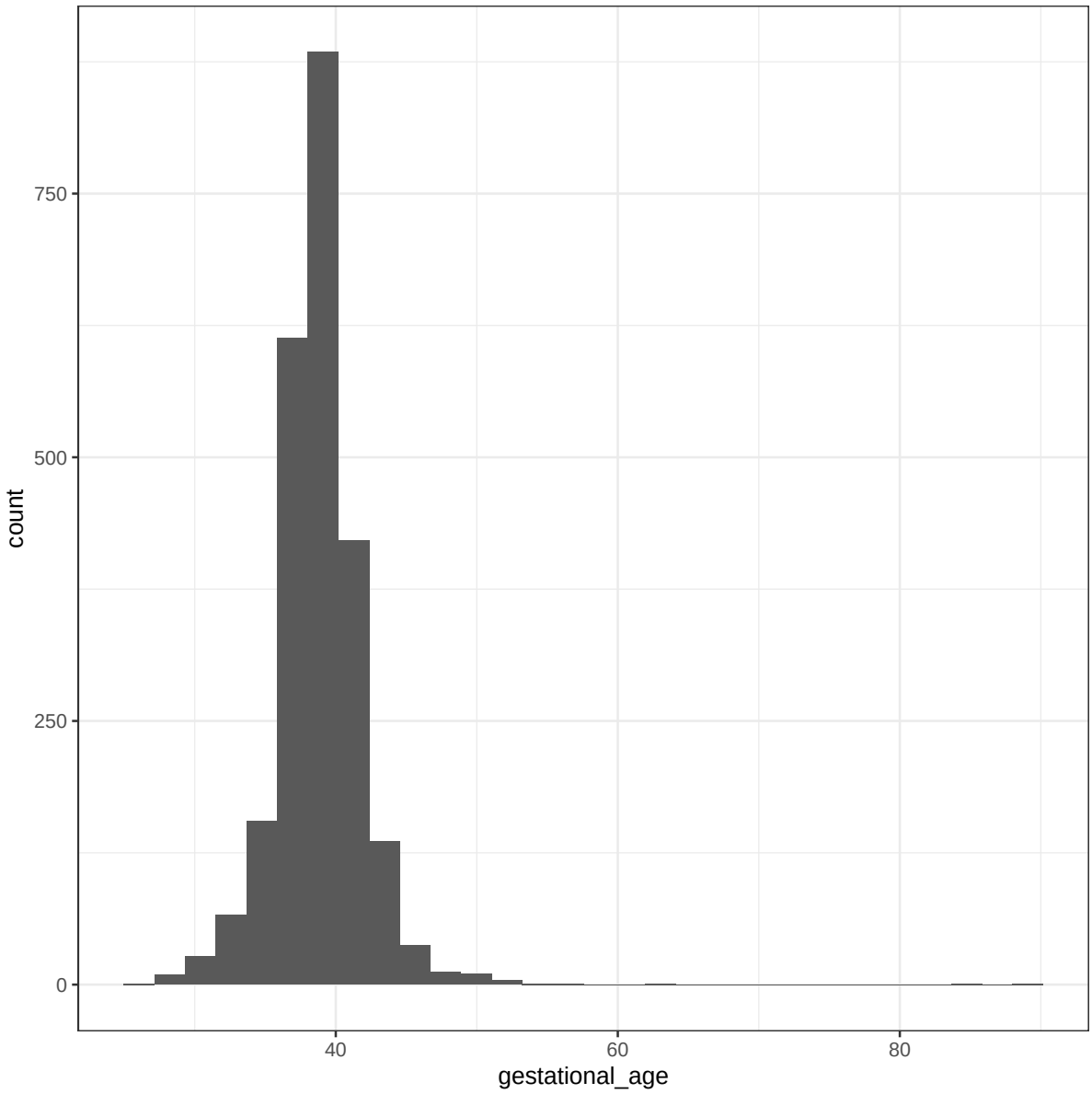
Exercise 2 EDA

We import the data and make a summary of the main values

```
# Import the data and observe the missing values / wierd values
data = readRDS("Longnecker.rds")
print(dim(data))

## [1] 2380   23

# The variables of interest are DDE, all the PBCs.
# The dependent variable is gestational_age
```

The plot of gestational age is

```
ggplot(data = data) +
  geom_histogram(aes(x=gestational_age))

## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```



We see in particular thea there are wierd values in the dependent variable. Indeed, the highest number of weeks obsrved is 90

```
table(data$gestational_age)
```

```
##
##  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46
##   1   4   5   8  19  32  34  67  88 103 161 349 425 459 279 142  79  57  25  12
##  47  48  49  50  51  52  53  55  56  64  84  90
##   8   4   4   3   3   3   1   1   1   1   1   1
```

As the record for longest pregnancy in weeks is equal to 375 and the second one is 317, we decide to drop all the observations which have a gestational age higher than 46 weeks (excluded). Moreoer, we transform smoking stauts and center as factors and we drop albumin (which is only made by NAs). Finally, we create a variable that reports if the the birth has been premature or not (before 37 weeks).

```
data = data %>%
  filter(gestational_age <= 46) %>%
```

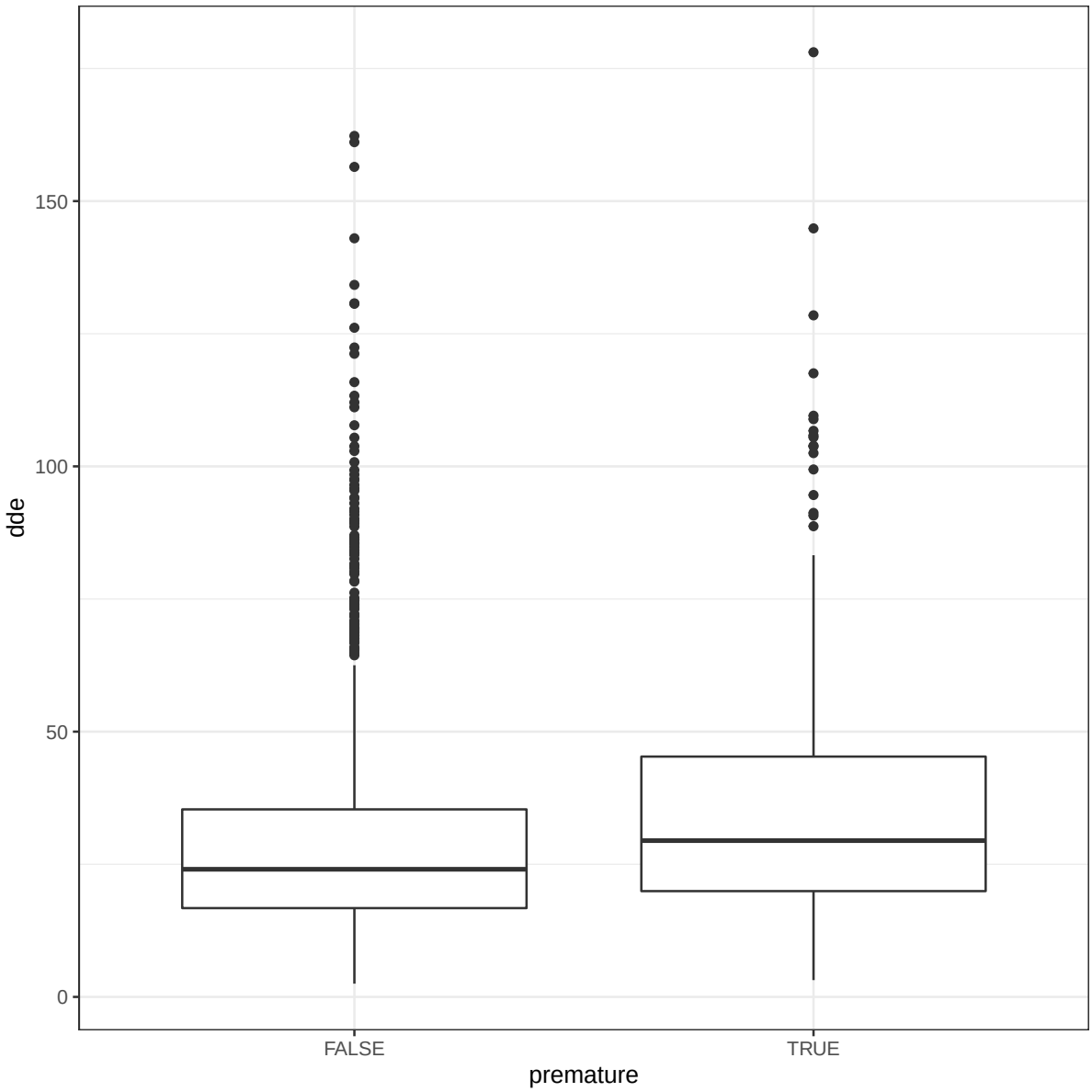
```
  mutate_at(vars(smoking_status, center),factor) %>%
  select(-albumin) %>% #Too many NAs
  mutate(premature = (gestational_age < 37))
```

We start by exploiting the relationship between the new variable premature and the DDE. Note that the number of premature mothers compared to the size of the new dataset is

```
table(data$premature)

##
## FALSE  TRUE
##  1988   361

table(data$premature)/nrow(data)

##
##     FALSE      TRUE
## 0.8463176 0.1536824
```

and the association it has with DDE is positive.

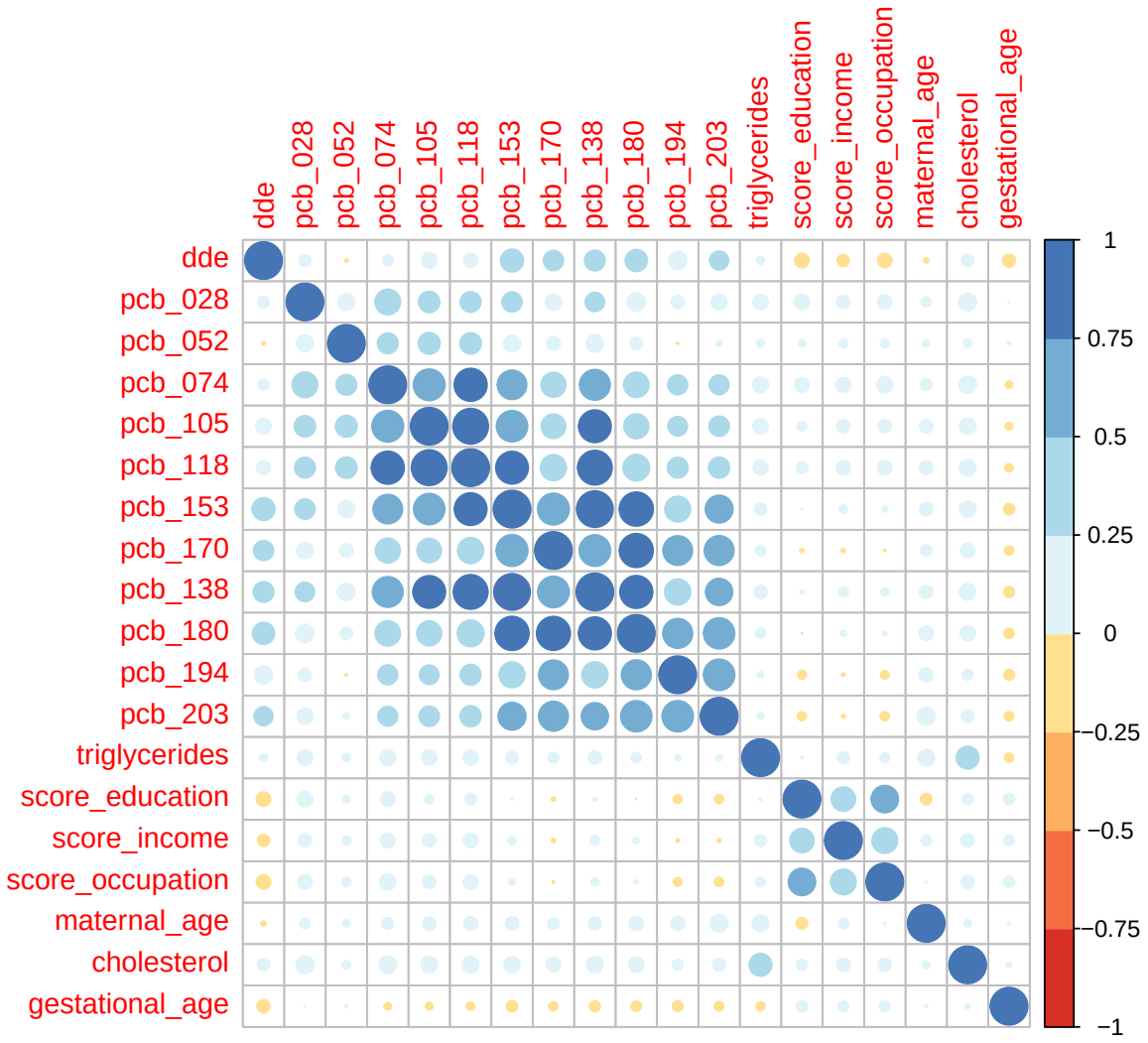
```
ggplot(data) +
  geom_boxplot(aes(x=premature, y=dde))
```



To see the relationship between premature and pcb, we have to understand what is the correlation across each level of pcb. Notice that there is only one row that has null values ofr pcb. We decide to drop it.

```
data = data[-is.na(data$pcb_028),]
```

```
cor_base = cor(data %>% select(-smoking_status, -race, -center, -premature), use="pairwise.complete.
corrplot(cor_base,col=brewer.pal(n=8, name="RdYlBu"))
```



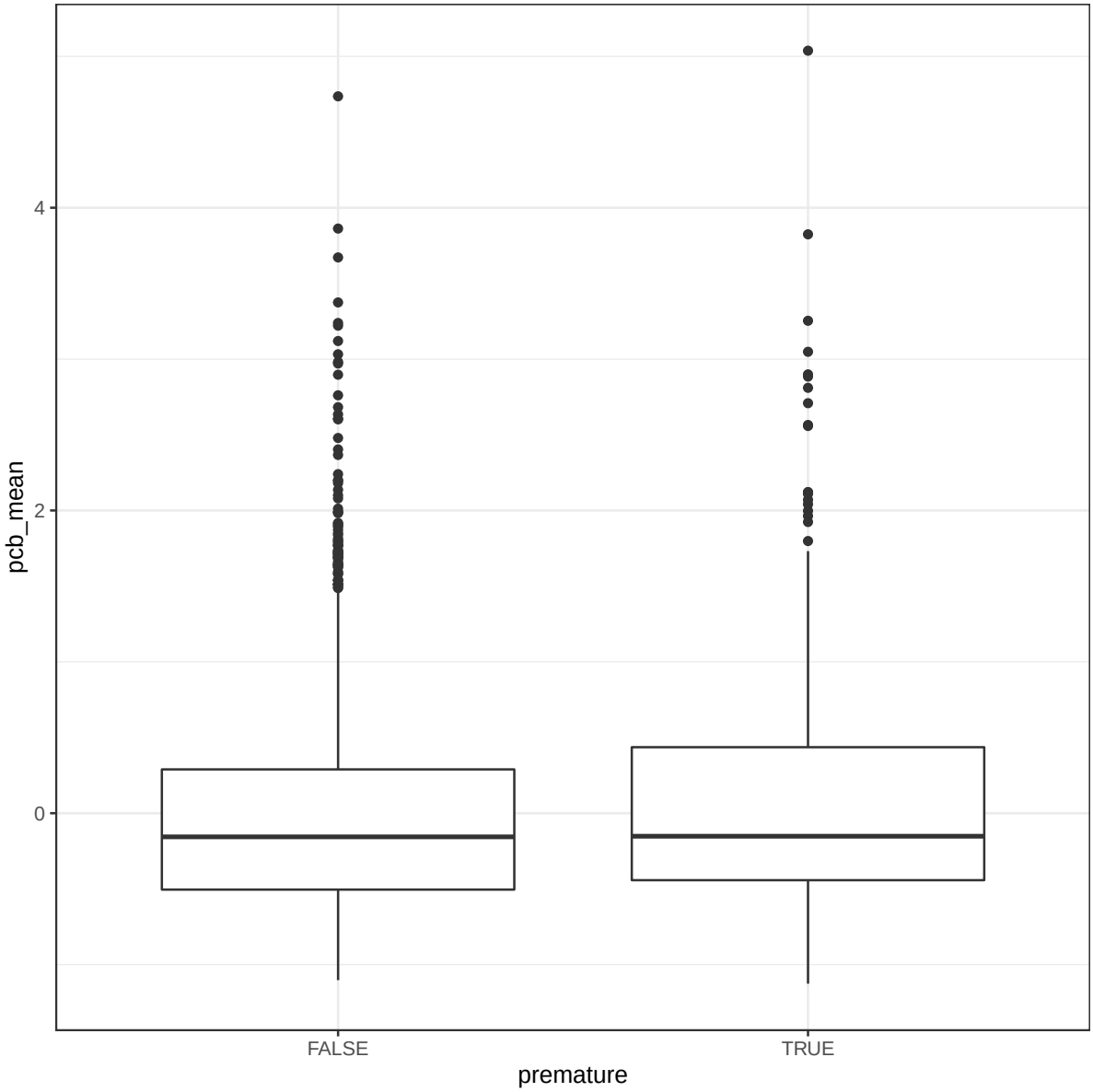
We see that each pcb are very correlated among themselves. This is expected, as they originate from one single component. As the variables have their own unit of measure, we suggest two approaches.

1. Summing them (to see the total amount of pcb in the blood).
2. Standardizing and averaging them

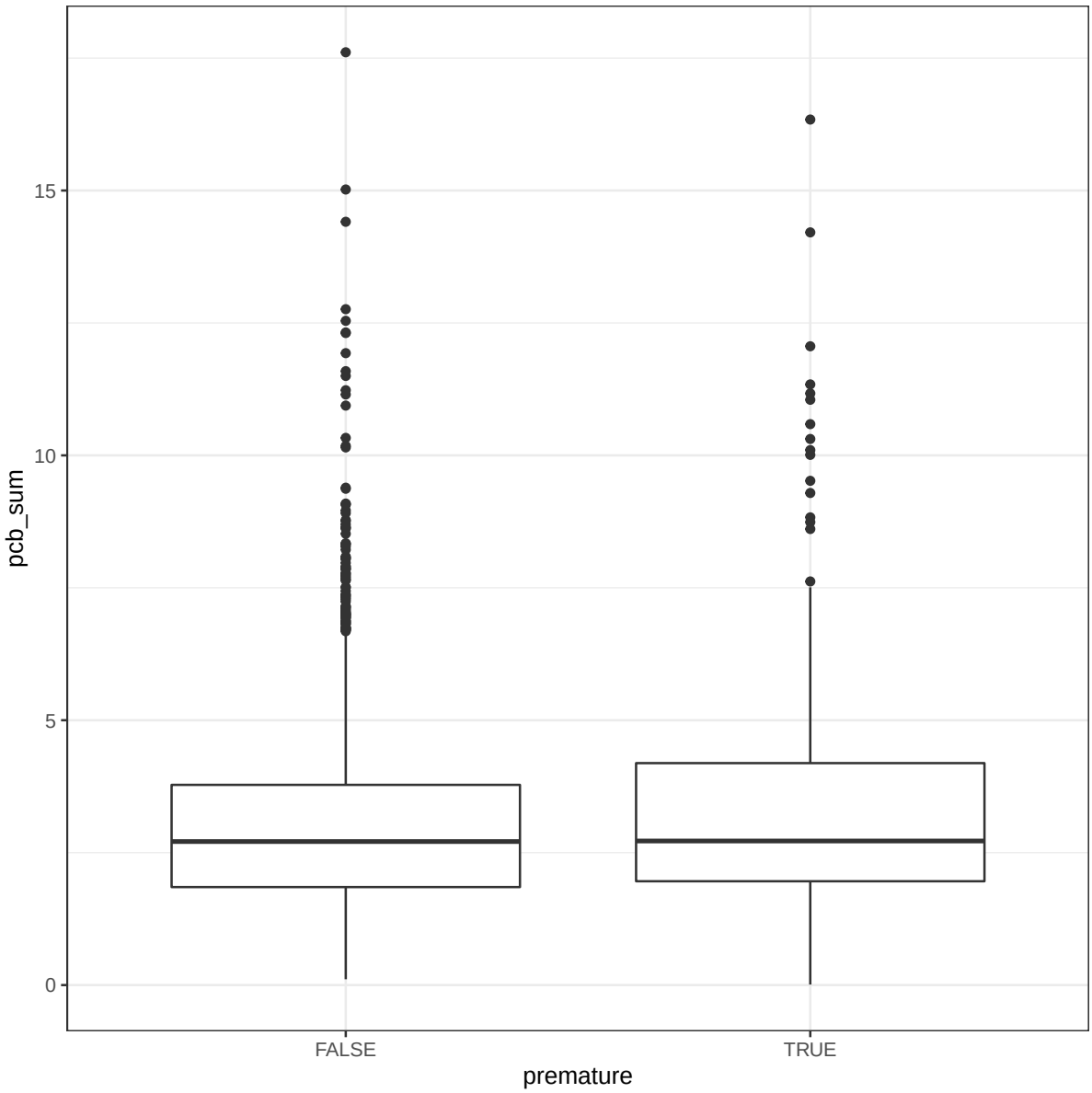
```
# 1) Summing
pcb_sum = apply(as.matrix(data %>% select(pcb_028, pcb_052, pcb_074, pcb_105, pcb_118, pcb_153, pcb_
# 2) Stanardize and average
my_standardize <- function(x) (x - mean(x, na.rm = T)) / sd(x, na.rm = T)
data = data %>%
  mutate_at(vars(starts_with("pcb")), my_standardize) %>% # standardize pcb's to give them all equal
  rowwise() %>%
  mutate(pcb_mean = mean(c(pcb_028, pcb_052, pcb_074, pcb_105, pcb_118, pcb_153, pcb_170, pcb_138, p
  ungroup
data$pcb_sum = pcb_sum
```

Thus, we can mirror the boxplot with DDE and premature also in this case.

```
ggplot(data) +
  geom_boxplot(aes(x=premature, y= pcb_mean))

## Warning:  Removed 1 rows containing non-finite values (stat_boxplot).
```

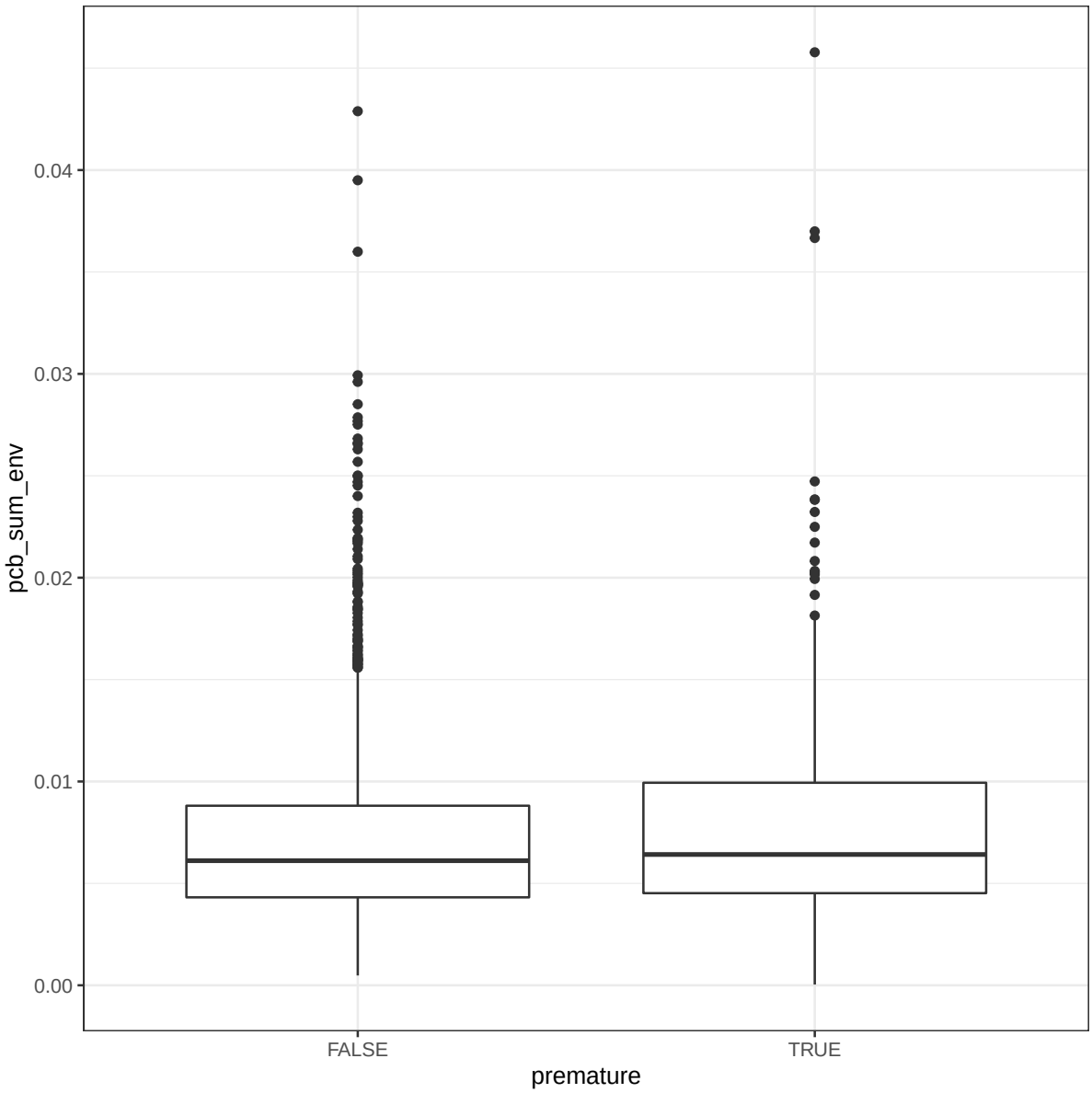



```
ggplot(data) +
  geom_boxplot(aes(x=premature, y= pcb_sum))

## Warning:  Removed 1 rows containing non-finite values (stat_boxplot).
```



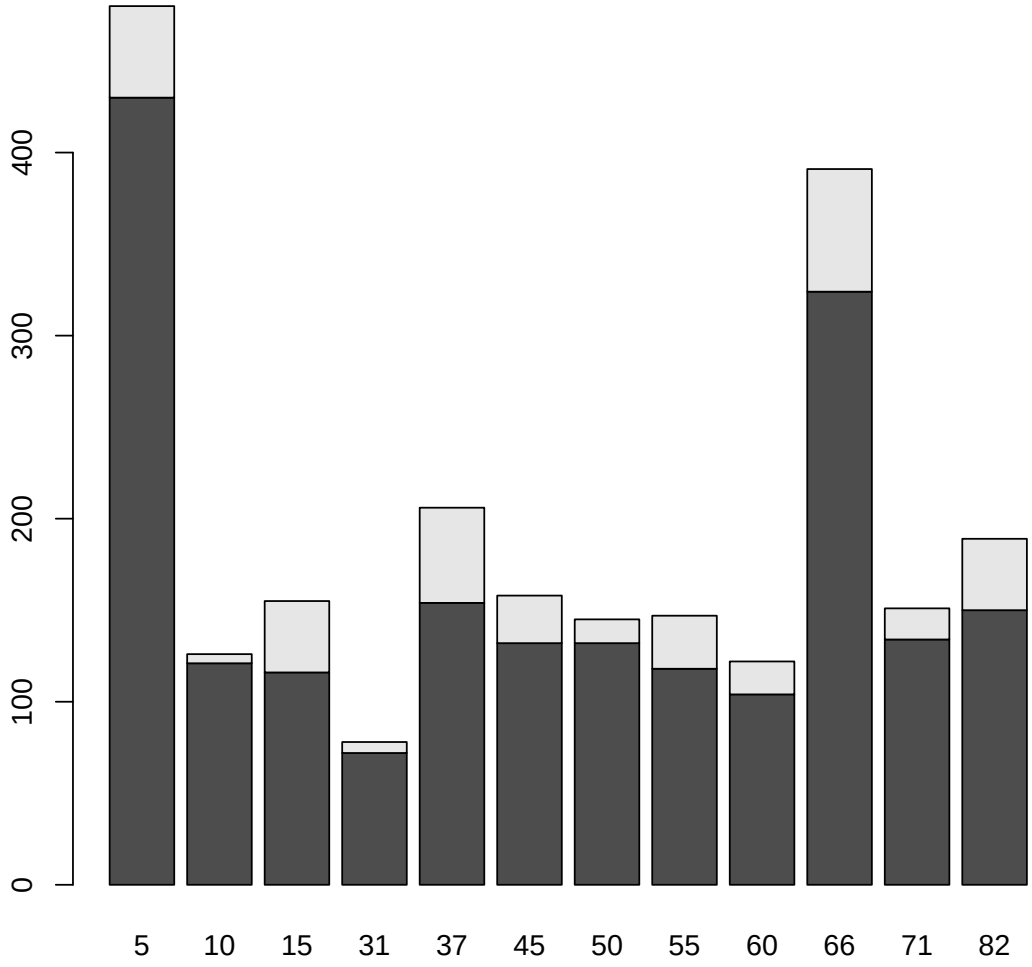
We see that the results are very similar. This suggests that the approaches will lead to the same results. At last, we isolate the results between environment and not environment (as raphael suggested).

```
cor(data %>% select(triglycerides, cholesterol))

##               triglycerides cholesterol
## triglycerides     1.0000000   0.3610952
## cholesterol       0.3610952   1.0000000

data = data %>% rowwise() %>% mutate(trigl_chol_sum = sum(triglycerides, cholesterol)) %>% ungroup
data$dde_env = data$dde/data$trigl_chol_sum
data$pcb_sum_env = data$pcb_sum/data$trigl_chol_sum
ggplot(data) +
  geom_boxplot(aes(x=premature, y= pcb_sum_env))

## Warning:  Removed 1 rows containing non-finite values (stat_boxplot).
```



One last thing. How are the premature women distributed across the centers?

```
data$premature = as.numeric(data$premature)
barplot(table(data$premature, data$center))
```



From a first glance,the labs seem eterogeneous. This is a problem we need to deal with.

Exercise 3 A first approach: logistic regression

We run a very simple logistic regression. First, we focus on the complete cases (to add all regressors). We start in the simple case of the dde and pcb$_s$*umwithouttheenvironmenteffect*.

```
logit_model_1 <- glm(premature ~ dde + pcb_sum + race + maternal_age + score_occupation + center +
summary(logit_model_1)

##
## Call:
## glm(formula = premature ~ dde + pcb_sum + race + maternal_age +
##     score_occupation + center + score_income + score_education,
##     family = "binomial", data = data)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3080  -0.6143  -0.4804  -0.3578   2.7137
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.116247   0.421263  -5.024 5.07e-07 ***
## dde               0.009322   0.003254   2.865  0.00417 **
## pcb_sum           0.107112   0.039410   2.718  0.00657 **
## raceblack         0.258101   0.241582   1.068  0.28535
## raceother         0.803932   0.494812   1.625  0.10422
## maternal_age     -0.004595   0.011105  -0.414  0.67906
## score_occupation -0.003767   0.002830  -1.331  0.18322
## center10         -0.870941   0.495521  -1.758  0.07881 .
## center15          0.580670   0.371618   1.563  0.11816
## center31         -1.003980   0.592436  -1.695  0.09014 .
## center37          0.652792   0.303428   2.151  0.03145 *
## center45         -0.028248   0.357586  -0.079  0.93704
## center50          0.032301   0.367842   0.088  0.93003
## center55         -0.448572   0.599546  -0.748  0.45435
## center60          0.251547   0.366731   0.686  0.49276
## center66          0.024323   0.314908   0.077  0.93843
## center71         -0.137112   0.362607  -0.378  0.70533
## center82          0.325151   0.367403   0.885  0.37616
## score_income     -0.002296   0.002782  -0.825  0.40934
## score_education  -0.004091   0.003124  -1.310  0.19036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1549.5  on 1831  degrees of freedom
## Residual deviance: 1453.4  on 1812  degrees of freedom
##   (516 observations deleted due to missingness)
## AIC: 1493.4
##
## Number of Fisher Scoring iterations: 5
```

```
logit_model_2 <- glm(premature ~ dde_env + pcb_sum_env + race + maternal_age + score_occupation + ce
summary(logit_model_2)

##
## Call:
## glm(formula = premature ~ dde_env + pcb_sum_env + race + maternal_age +
##     score_occupation + center + score_income + score_education,
##     family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0686  -0.6225  -0.4909  -0.3649   2.6527
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.922867   0.418335  -4.596  4.3e-06 ***
```

```
## dde_env           3.042856   1.353676   2.248   0.0246 *
## pcb_sum_env      20.220561  16.410417   1.232   0.2179
## raceblack         0.279208   0.244582   1.142   0.2536
## raceother         0.719970   0.491241   1.466   0.1428
## maternal_age      0.000080   0.011020   0.007   0.9942
## score_occupation -0.003359   0.002816  -1.193   0.2330
## center10         -0.821461   0.492568  -1.668   0.0954 .
## center15          0.431834   0.368169   1.173   0.2408
## center31         -0.971114   0.587981  -1.652   0.0986 .
## center37          0.607774   0.302309   2.010   0.0444 *
## center45         -0.029969   0.354556  -0.085   0.9326
## center50         -0.094126   0.364143  -0.258   0.7960
## center55         -0.507646   0.593630  -0.855   0.3925
## center60          0.129617   0.363120   0.357   0.7211
## center66         -0.019042   0.312753  -0.061   0.9515
## center71         -0.195226   0.360729  -0.541   0.5884
## center82          0.191025   0.365205   0.523   0.6009
## score_income     -0.002060   0.002756  -0.748   0.4547
## score_education  -0.004532   0.003105  -1.460   0.1444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1549.5  on 1831  degrees of freedom
## Residual deviance: 1466.5  on 1812  degrees of freedom
##   (516 observations deleted due to missingness)
## AIC: 1506.5
##
## Number of Fisher Scoring iterations: 5
```