EDA

Raphaël Morsomme

January 15, 2020

```
library(tidyverse)
library(GGally) # ggpairs()

my_standardize <- function(x) (x - mean(x, na.rm = T)) / sd(x, na.rm = T)

# longest pregnancy possible
# http://content.time.com/time/magazine/article/0,9171,797153,00.html
(375-58)/7 # 45.28
```

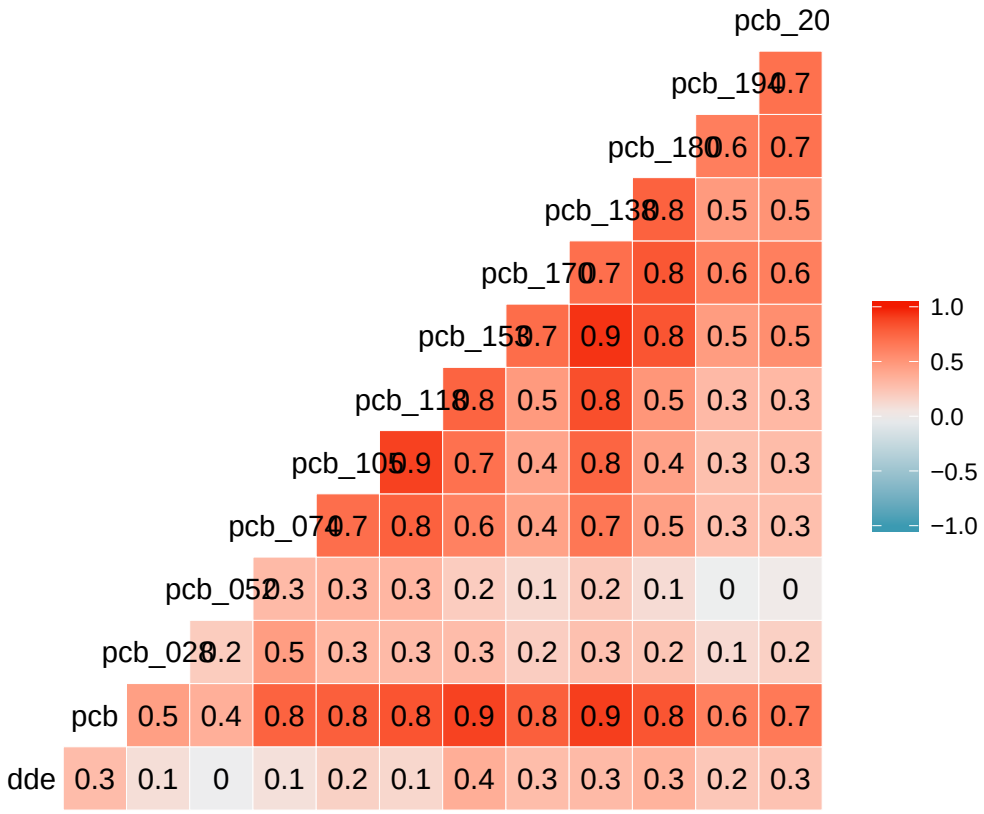
```
## [1] 45.28571
```

```
d <- readRDS("Longnecker.rds") %>%

  mutate_at(vars(center, smoking_status), factor) %>%
  select(-albumin) %>% # too many NAs

  filter(gestational_age <= 46) %>% # more accurate treatment would be to allow for error in measuremen
  mutate(pre_mature = gestational_age <= 37) %>%


  # construct aggregate pcb variable: average of pcb's
  mutate_at(vars(starts_with("pcb")), my_standardize) %>% # standardize pcb's to give them all equal we
  rowwise() %>%
  mutate(pcb = mean(c(pcb_028, pcb_052, pcb_074, pcb_105, pcb_118, pcb_153, pcb_170, pcb_138, pcb_180, p
  ungroup
```

```
sum(complete.cases(d)) / nrow(d)
```

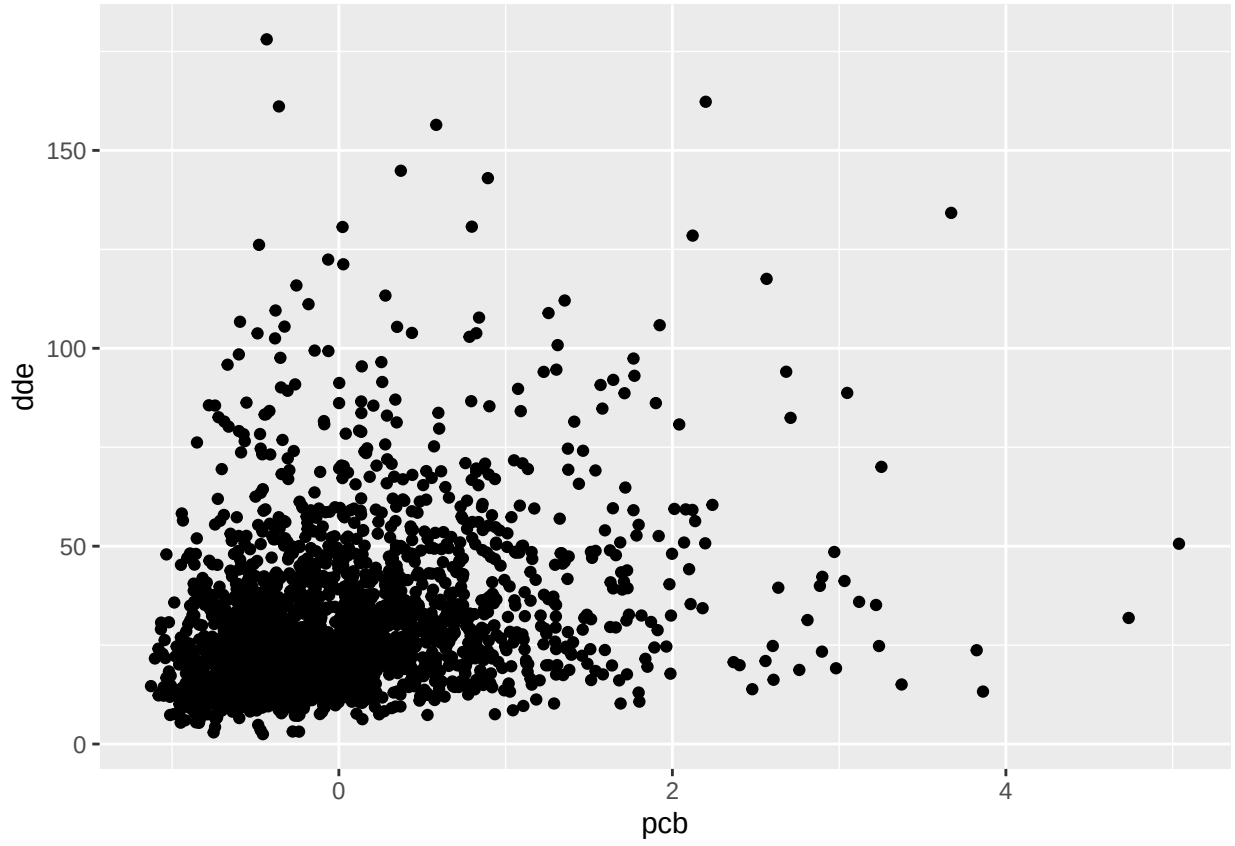
```
## [1] 0.7803321
```

```
d %>%
  select(dde, pcb, pcb_028 : pcb_203) %>%
  #mutate_all(~(log(. + 0.1))) %>%
  ggcorr(palette = "RdBu", label = TRUE)
```



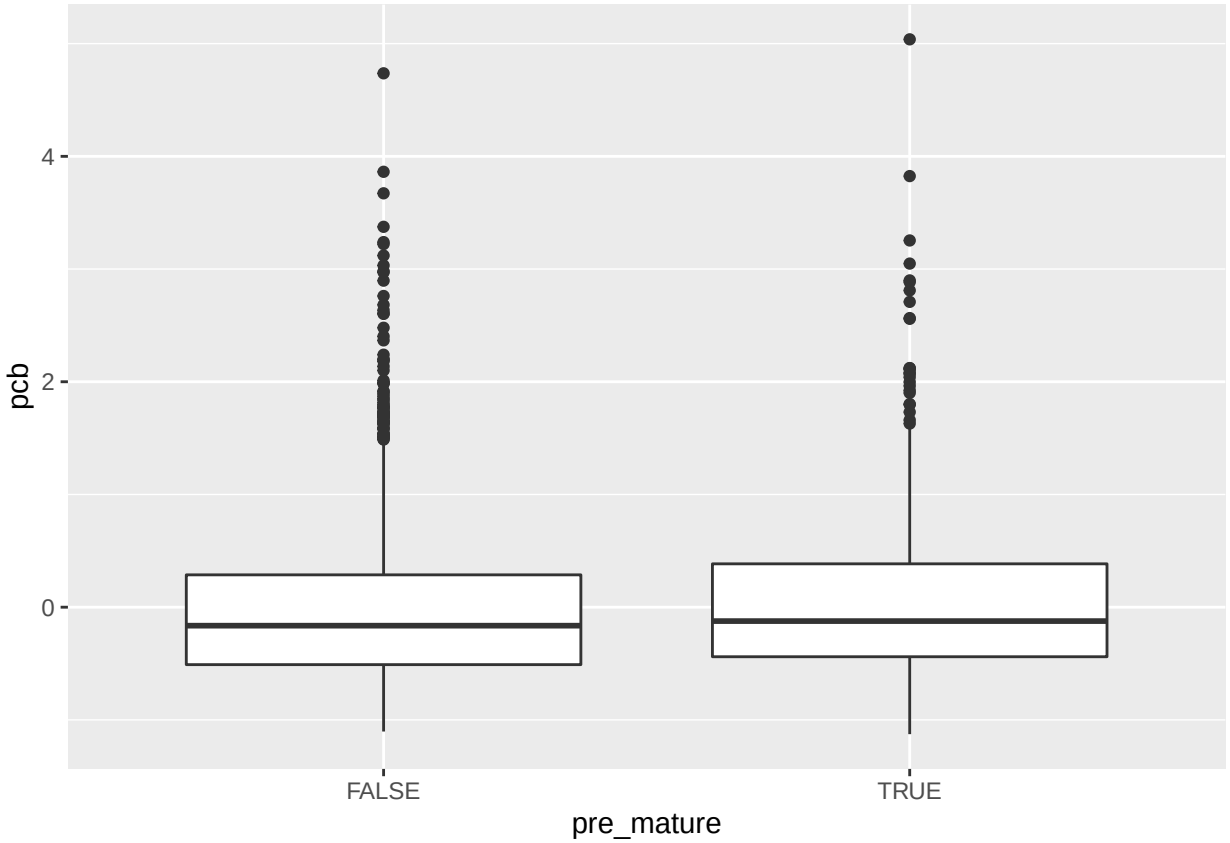

```
d %>%
  ggplot(aes(x = pcb, y = dde)) +
  geom_point()
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



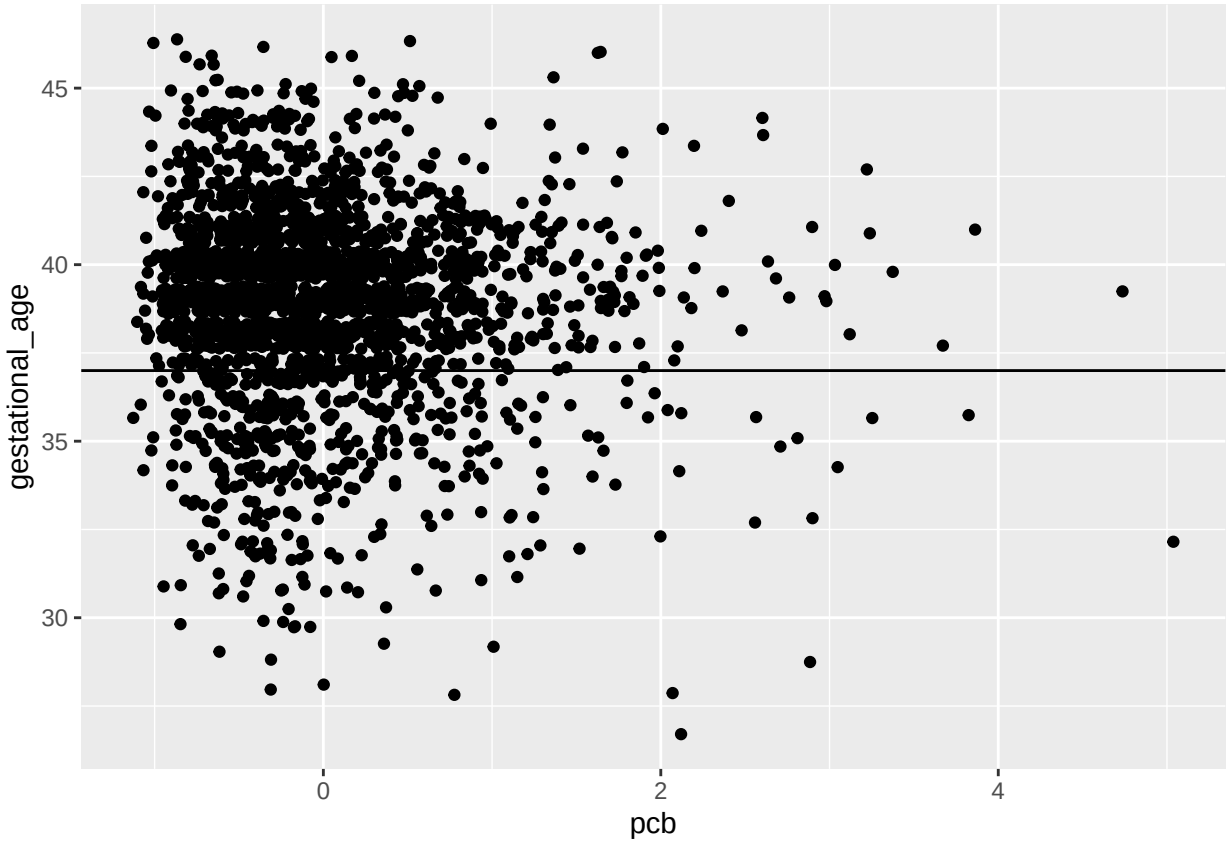

```
d %>%
  ggplot(aes(x = pre_mature, y = pcb)) +
  geom_boxplot()
```

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

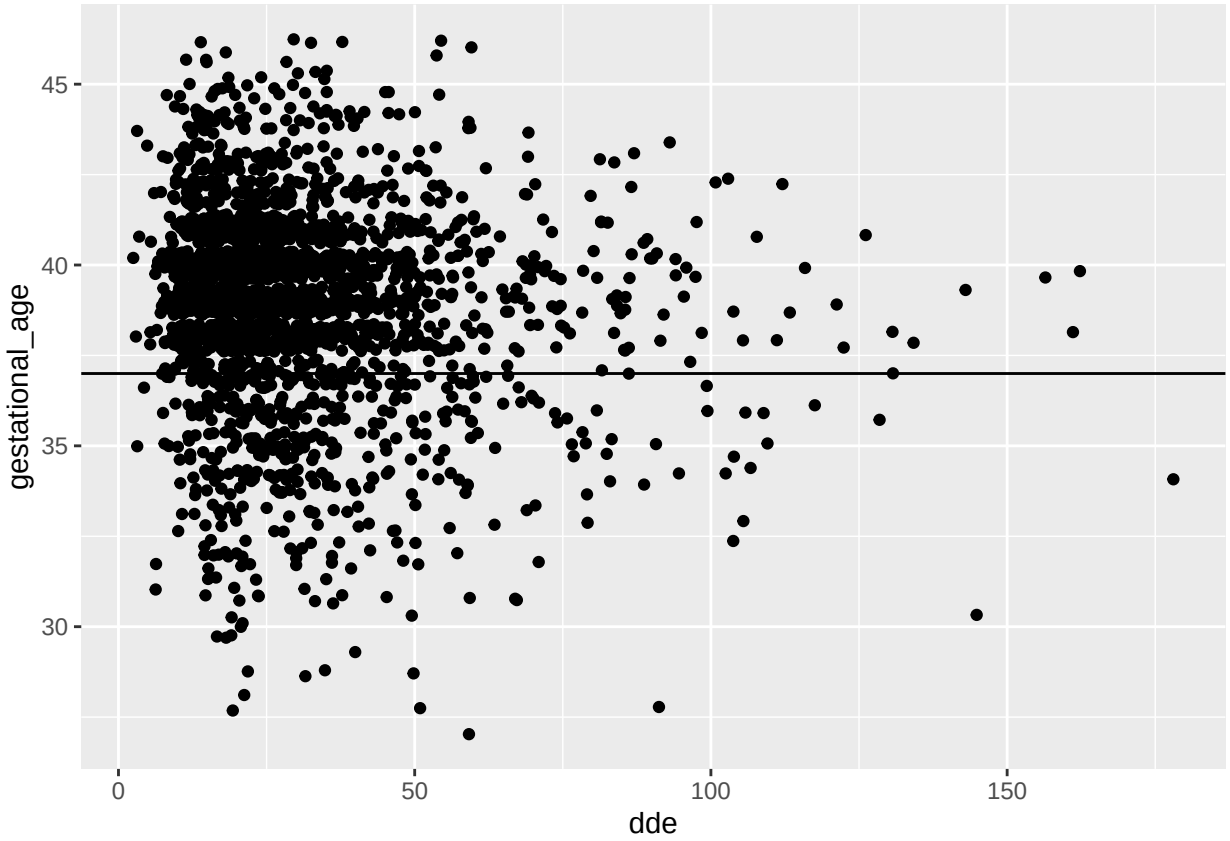



```
d %>%
  ggplot(aes(x = pcb, y = gestational_age)) +
  geom_jitter() +
  geom_hline(yintercept = 37)
```

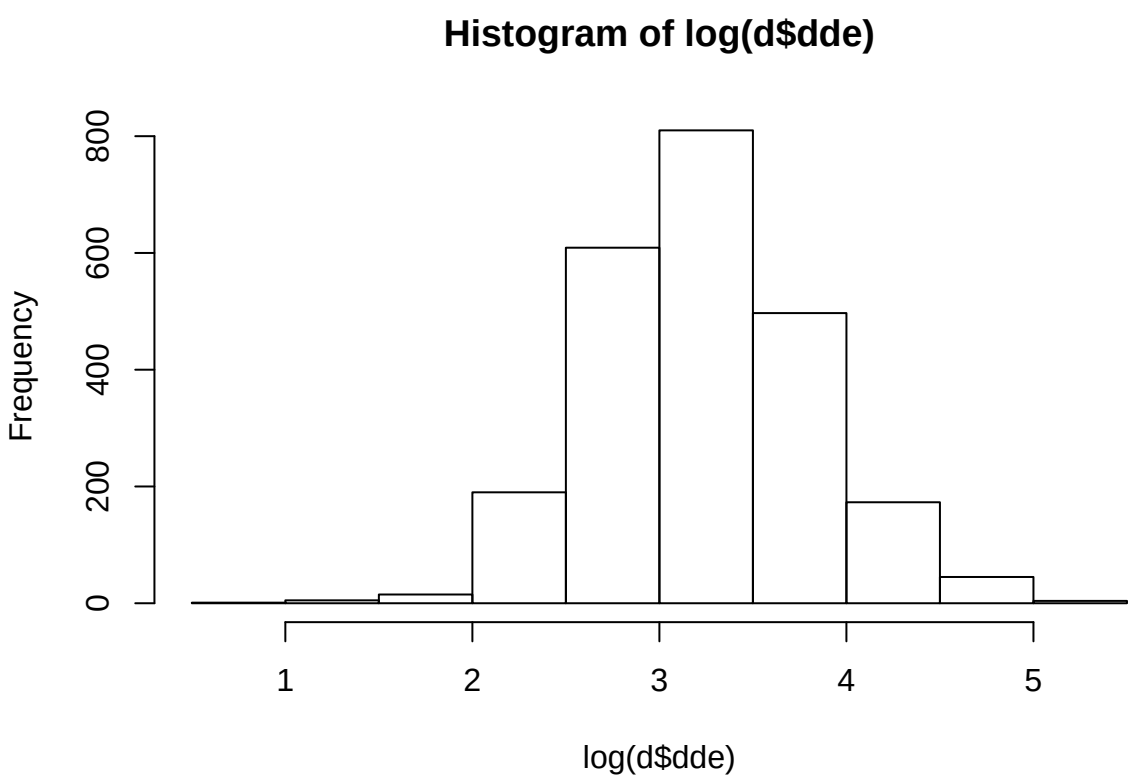
```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
d %>%
  ggplot(aes(x = dde, y = gestational_age)) +
  geom_jitter() +
  geom_hline(yintercept = 37)
```

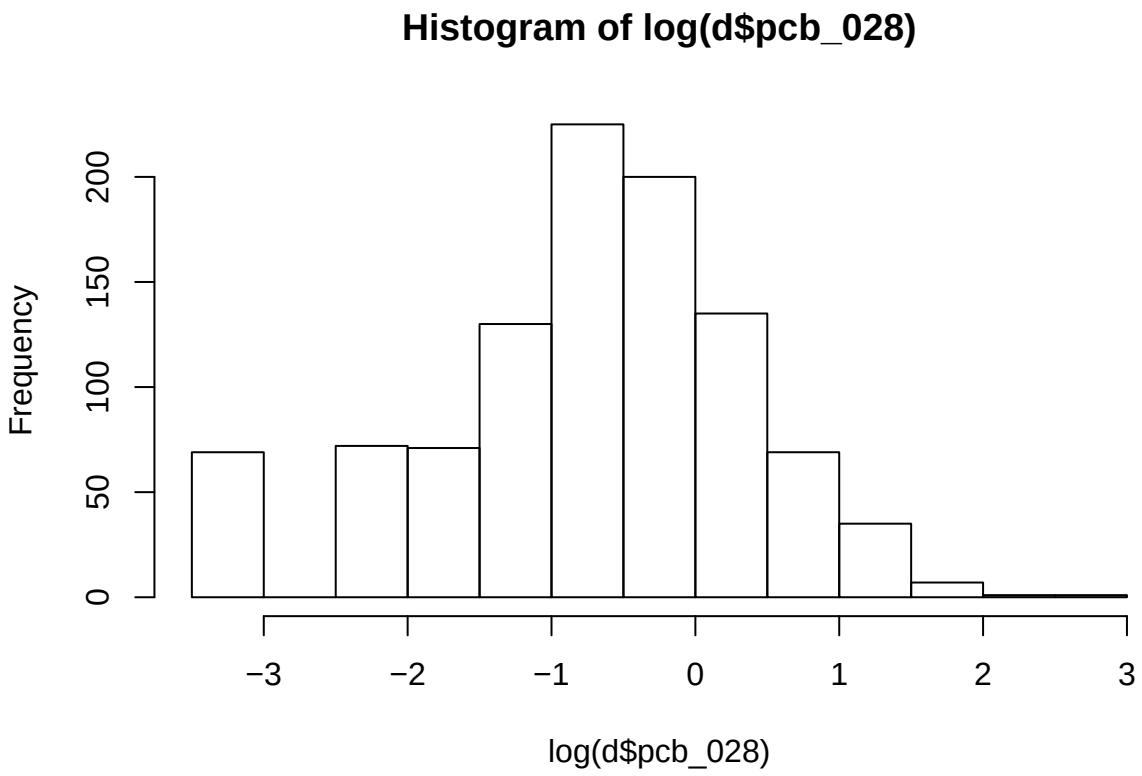



```
hist(log(d$dde))
```



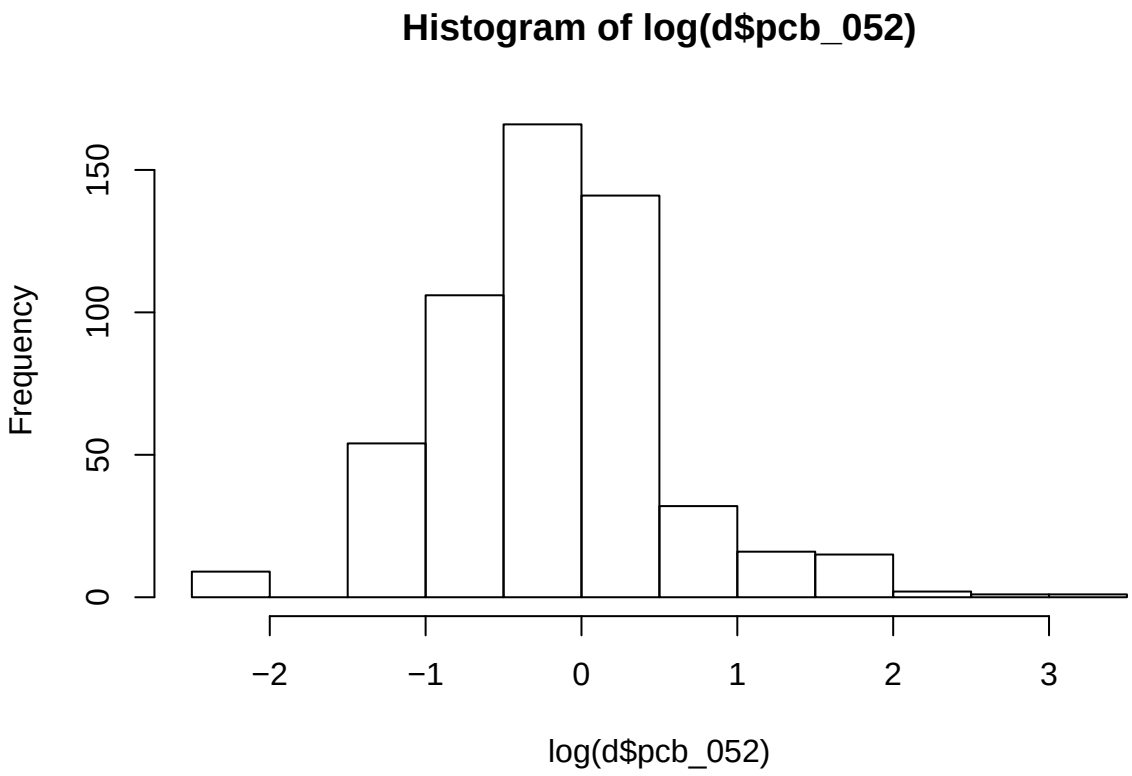

```
hist(log(d$pcb_028))
```

```
## Warning in log(d$pcb_028): NaNs produced
```

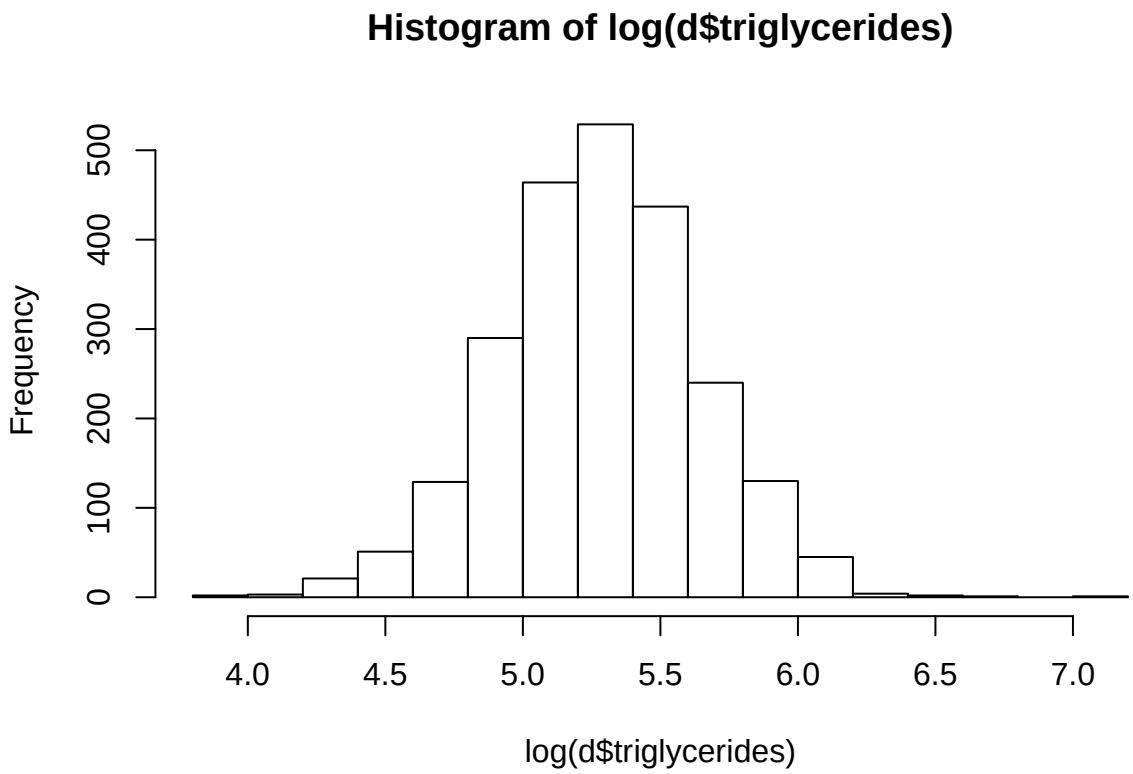



```
hist(log(d$pcb_052))
```

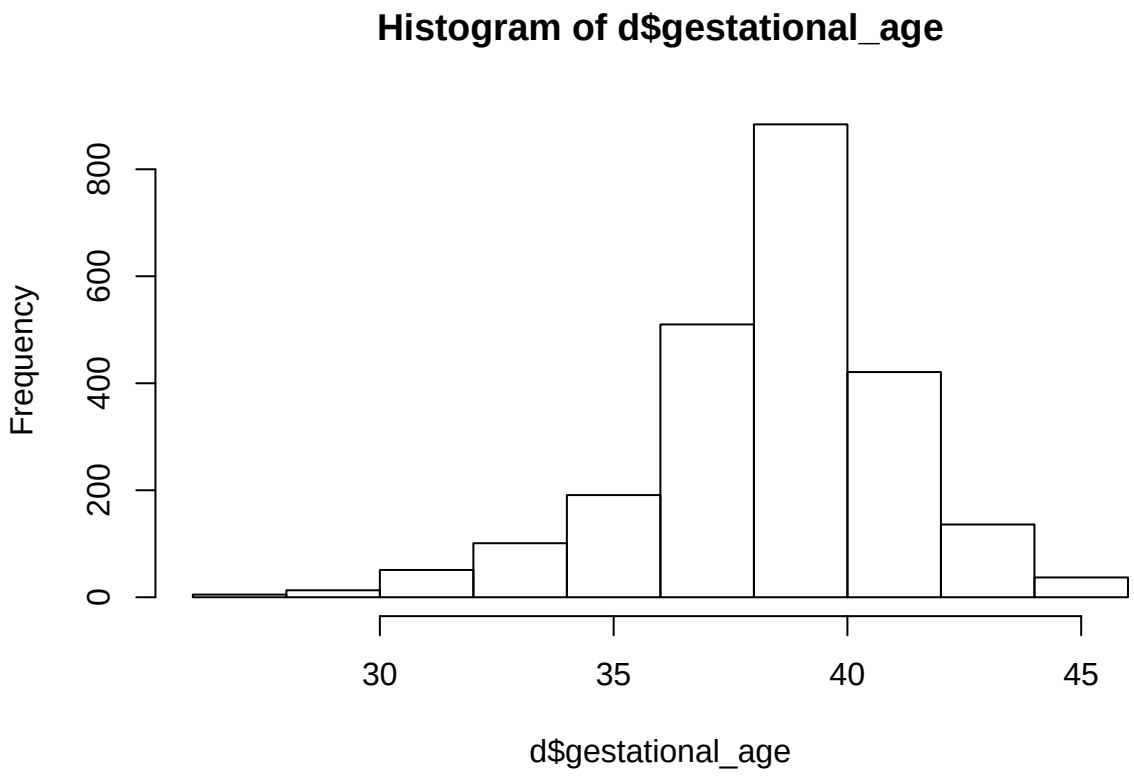
```
## Warning in log(d$pcb_052): NaNs produced
```




```
hist(log(d$triglycerides))
```




```
hist(d$gestational_age)
```




```
sort(d$gestational_age, T)[1:10]
```

```
##  [1] 46 46 46 46 46 46 46 46 46 46
```

```
hist(d$maternal_age)
```


```
d %>%
  select(cholesterol, triglycerides, pcb, dde) %>%
  ggpairs(lower = list(continuous = wrap("points", alpha = 0.5, size = .1)))
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removing 1 row that contained a missing value
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removing 1 row that contained a missing value
```

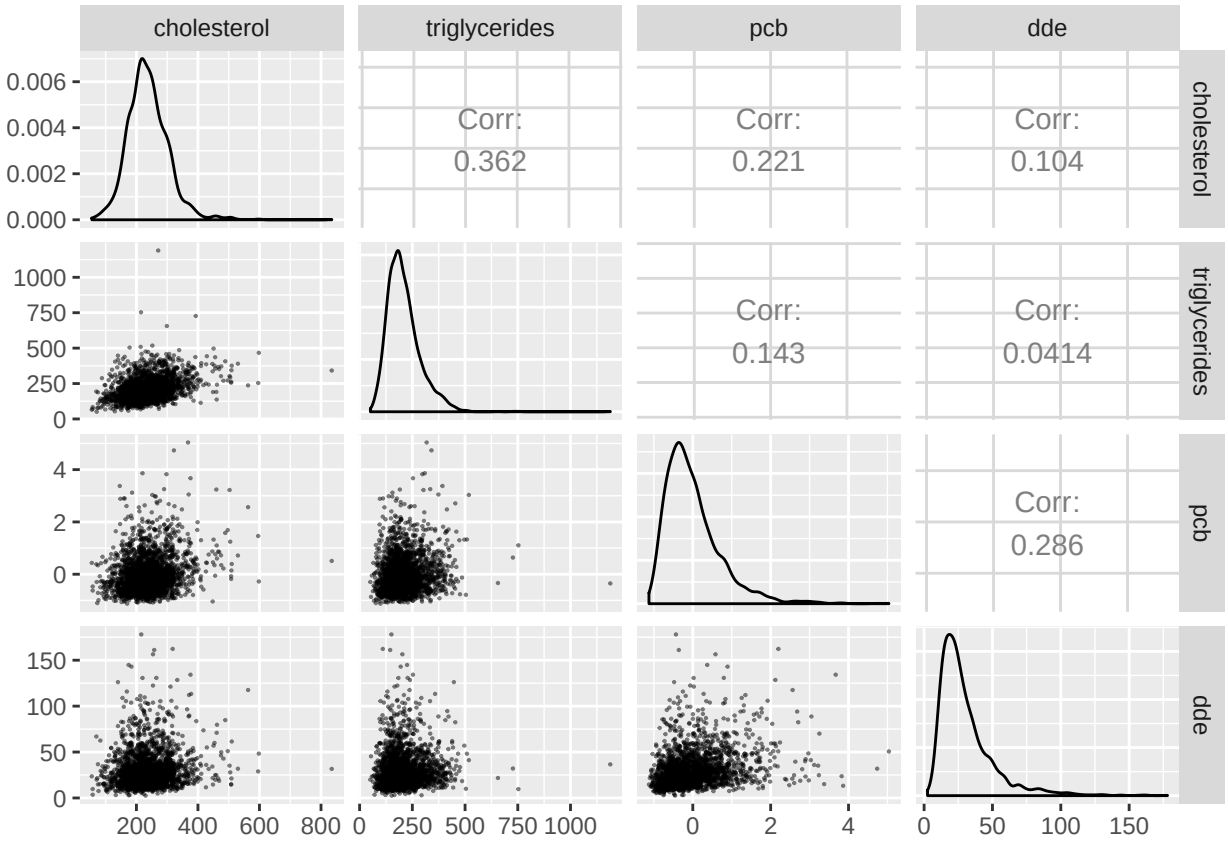
```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

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

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removing 1 row that contained a missing value
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```


cholesterol
triglycerides
pcb
dde
Corr:
0.362
Corr:
0.221
Corr:
0.104
Corr:
0.143
Corr:
0.0414
Corr:
0.286
0.006
0.004
0.002
0.000
1000
750
500
250
0
4
2
0
150
100
50
0
200
400
600
800
0
250
500
750
1000
0
2
4
0
50
100
150