"exploration data analysis of CpG methylation data"

Kim Reijntjens

20-9-2022

relevance of the project

CpG sites are often described in the study called epigenetics: "where genetic expression is not the direct result of the information stored in the nucleotide sequence of DNA. Instead, the DNA is altered in a way that affects its expression. These changes are stable in the sense that they are transmitted during cell division to progeny cells, and often through gametes to future generations. The precise molecular mechanism of imprinting and other epigenetic events is still a matter for conjecture, but it seems certain that DNA methylation is involved. In most eukaryotes, methyl groups can be added to the carbon atom at position 5 in cytosine (see Chapter 10) as a result of the activity of the enzyme DNA methyltransferase. Methyl groups are added when the dinucleotide CpG or groups of CpG units (called CpG islands) are present along a DNA chain. DNA methylation is a reasonable mechanism for establishing a molecular imprint, since there is evidence that a high level of methylation can inhibit gene activity and that active genes (or their regulatory sequences) are often undermethylated." (Klug, William S.;Cummings, Michael R.;Spencer, Charlotte A.;Michael A. Palladino - Concepts of genetics)

because alterations in the genome and activity of the genes are associated with common diseases such as cancer or asthma, it is important to know if smoking can be the cause of this. If this is the case than in further research there can be looked at which CpG sites are affected by smoking, and if there is a relation with the affected CpG sites and the ones associated with diseases like cancer and asthma. (Klug, William S.;Cummings, Michael R.;Spencer, Charlotte A.;Michael A. Palladino - Concepts of genetics)

goal

can CpG methylation show a relation with smoking, based on a prediction whether a patient is smoking or non-smoking using differences in CpG site values.

For this we use the dataset of 683 patients. the dataset has 683 rows and 24 columns containing the patients; age, gender, smoking status and 20 CpG site values. the dataset was already compressed to 20 CpG site values sites per sample. The original dataset was composed for a study titled "Differential DNA methylation in Rheumatoid arthritis" (NCBI series GSE42861) Where the original dataset contained 485577 rows with methylation data per sample (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42861) (https://www.kaggle.com/datasets/thomaskonstantin/cpg-values-of-smoking-and-non-smoking-patients)

exploration of the data

```
library(pander)
patient_data <- read.csv(file = "data/Smoker_Epigenetic_df.csv")
head(patient_data)
```

```
##          GSM Smoking.Status Gender Age cg00050873 cg00212031 cg00213748
## 1 GSM1051525        current      f  67  0.6075634  0.4228427  0.3724549
## 2 GSM1051526        current      f  49  0.3450542  0.5686615  0.5005995
## 3 GSM1051527        current      f  53  0.3213497  0.3609091  0.3527315
## 4 GSM1051528        current      f  62  0.2772675  0.3044371  0.4752352
## 5 GSM1051529          never      f  33  0.4135991  0.1312511  0.3675446
## 6 GSM1051530        current      f  59  0.6228599  0.5016849  0.2632270
##   cg00214611 cg00455876 cg01707559 cg02004872 cg02011394 cg02050847 cg02233190
## 1  0.6215619  0.2907773  0.2671431  0.1791439  0.4802517  0.3276078  0.2411204
## 2  0.4986067  0.3745909  0.1902743  0.1559775  0.4180809  0.3464627  0.1754907
## 3  0.3738240  0.2306740  0.3147052  0.1057448  0.6151030  0.2375392  0.2464092
## 4  0.4862581  0.2951815  0.2957931  0.1112862  0.3010196  0.3045353  0.1770279
## 5  0.7611667  0.2357703  0.2505265  0.1691084  0.3929746  0.3062257  0.3017014
## 6  0.4157459  0.4751891  0.2539041  0.2607587  0.5097921  0.4052457  0.3852716
##   cg02494853  cg02839557 cg02842889 cg03052502 cg03155755 cg03244189 cg03443143
## 1 0.06706958 0.246993368  0.4692396  0.4002466  0.4150313  0.2214331  0.4758258
## 2 0.04693889 0.236742313  0.3074666  0.3770313  0.3973715  0.2171221  0.5444690
## 3 0.03823712 0.244611725  0.3577526  0.3050442  0.5212775  0.1850495  0.5370600
## 4 0.02671625 0.001641439  0.4457390  0.2714746  0.4344920  0.1654187  0.5079167
## 5 0.03701636 0.334319727  0.3950396  0.3265530  0.4300966  0.1811352  0.4054791
## 6 0.02583463 0.309210202  0.3218573  0.5333670  0.5715522  0.2109749  0.3778239
##   cg03683899 cg03695421 cg03706273
## 1  0.2077242  0.2091974 0.12998255
## 2  0.1844462  0.1937732 0.09853265
## 3  0.3931231  0.2680030 0.04024808
## 4  0.2812089  0.2178572 0.10151626
## 5  0.3107944  0.2800708 0.07785712
## 6  0.4693609  0.3433317 0.04577912
```

```
#View(patient_data)


pander::pander(summary(patient_data))
```

Table 1: Table continues below

GSM	Smoking.Status	Gender	Age
Length:683	Length:683	Length:683	Min. :18.00
Class :character	Class :character	Class :character	1st Qu.:47.00
Mode :character	Mode :character	Mode :character	Median :56.00
NA	NA	NA	Mean :53.82
NA	NA	NA	3rd Qu.:62.00
NA	NA	NA	Max. :80.00
NA	NA	NA	NA

Table 2: Table continues below

cg00050873	cg00212031	cg00213748	cg00214611
Min. :0.1186	Min. :0.00695	Min. :0.0000	Min. :0.01247
1st Qu.:0.4131	1st Qu.:0.06317	1st Qu.:0.3635	1st Qu.:0.06946
Median :0.5052	Median :0.36554	Median :0.4713	Median :0.41575

cg00050873	cg00212031	cg00213748	cg00214611
Mean :0.5600	Mean :0.30960	Mean :0.5191	Mean :0.34106
3rd Qu.:0.8144	3rd Qu.:0.45981	3rd Qu.:0.7278	3rd Qu.:0.49745
Max. :0.8989	Max. :0.70999	Max. :0.9236	Max. :0.80606
NA's :62	NA's :62	NA's :62	NA's :62

Table 3: Table continues below

cg00455876	cg01707559	cg02004872	cg02011394
Min. :0.05917	Min. :0.04333	Min. :0.00262	Min. :0.0000
1st Qu.:0.29300	1st Qu.:0.11080	1st Qu.:0.04284	1st Qu.:0.4261
Median :0.37968	Median :0.23873	Median :0.14933	Median :0.5157
Mean :0.44718	Mean :0.21435	Mean :0.15542	Mean :0.6058
3rd Qu.:0.66283	3rd Qu.:0.28061	3rd Qu.:0.24263	3rd Qu.:0.9412
Max. :0.85443	Max. :0.46999	Max. :0.47384	Max. :0.9792
NA's :62	NA's :62	NA's :62	NA's :62

Table 4: Table continues below

cg02050847	cg02233190	cg02494853	cg02839557
Min. :0.05234	Min. :0.00863	Min. :0.01162	Min. :0.00000
1st Qu.:0.33963	1st Qu.:0.08838	1st Qu.:0.02865	1st Qu.:0.06384
Median :0.42754	Median :0.25982	Median :0.03695	Median :0.35042
Mean :0.54369	Mean :0.23250	Mean :0.04077	Mean :0.30088
3rd Qu.:0.95558	3rd Qu.:0.33702	3rd Qu.:0.04677	3rd Qu.:0.45786
Max. :0.98320	Max. :0.51173	Max. :0.28947	Max. :0.82739
NA's :62	NA's :62	NA's :62	NA's :62

Table 5: Table continues below

cg02842889	cg03052502	cg03155755	cg03244189
Min. :0.01346	Min. :0.0000	Min. :0.2020	Min. :0.02972
1st Qu.:0.05483	1st Qu.:0.4025	1st Qu.:0.4245	1st Qu.:0.11976
Median :0.39757	Median :0.4940	Median :0.4962	Median :0.20397
Mean :0.32362	Mean :0.5907	Mean :0.5895	Mean :0.19552
3rd Qu.:0.47385	3rd Qu.:0.9631	3rd Qu.:0.8988	3rd Qu.:0.24921
Max. :0.85625	Max. :0.9902	Max. :0.9696	Max. :0.54074
NA's :62	NA's :62	NA's :62	NA's :62

cg03443143	cg03683899	cg03695421	cg03706273
Min. :0.06496	Min. :0.00788	Min. :0.0949	Min. :0.01120
1st Qu.:0.40963	1st Qu.:0.06159	1st Qu.:0.2566	1st Qu.:0.03413
Median :0.48314	Median :0.34422	Median :0.3208	Median :0.04961
Mean :0.56841	Mean :0.28442	Mean :0.3978	Mean :0.05769
3rd Qu.:0.85436	3rd Qu.:0.41866	3rd Qu.:0.5965	3rd Qu.:0.06916
Max. :0.93589	Max. :0.65925	Max. :0.8433	Max. :0.34380
NA's :62	NA's :62	NA's :62	NA's :62

```
panderOptions("table.continues")
```

```
## [1] "Table continues below"
```

```
#panderOptions("table.continues.affix")

pander::pander(summary(data), caption = "Summary with basic statistics about the data colums")
```

Quitting from lines 43-62 (CpG_methylation_log.Rmd) Error in object[[i]] : object of type 'closure' is not subsettable Calls: ... eval -> eval -> -> summary -> summary.default

```
str(patient_data)
```

```
## 'data.frame':    683 obs. of  24 variables:
##  $ GSM           : chr  "GSM1051525" "GSM1051526" "GSM1051527" "GSM1051528" ...
##  $ Smoking.Status: chr  "current" "current" "current" "current" ...
##  $ Gender        : chr  " f" " f" " f" " f" ...
##  $ Age           : int  67 49 53 62 33 59 66 51 55 37 ...
##  $ cg00050873    : num  0.608 0.345 0.321 0.277 0.414 ...
##  $ cg00212031    : num  0.423 0.569 0.361 0.304 0.131 ...
##  $ cg00213748    : num  0.372 0.501 0.353 0.475 0.368 ...
##  $ cg00214611    : num  0.622 0.499 0.374 0.486 0.761 ...
##  $ cg00455876    : num  0.291 0.375 0.231 0.295 0.236 ...
##  $ cg01707559    : num  0.267 0.19 0.315 0.296 0.251 ...
##  $ cg02004872    : num  0.179 0.156 0.106 0.111 0.169 ...
##  $ cg02011394    : num  0.48 0.418 0.615 0.301 0.393 ...
##  $ cg02050847    : num  0.328 0.346 0.238 0.305 0.306 ...
##  $ cg02233190    : num  0.241 0.175 0.246 0.177 0.302 ...
##  $ cg02494853    : num  0.0671 0.0469 0.0382 0.0267 0.037 ...
##  $ cg02839557    : num  0.24699 0.23674 0.24461 0.00164 0.33432 ...
##  $ cg02842889    : num  0.469 0.307 0.358 0.446 0.395 ...
##  $ cg03052502    : num  0.4 0.377 0.305 0.271 0.327 ...
##  $ cg03155755    : num  0.415 0.397 0.521 0.434 0.43 ...
##  $ cg03244189    : num  0.221 0.217 0.185 0.165 0.181 ...
##  $ cg03443143    : num  0.476 0.544 0.537 0.508 0.405 ...
##  $ cg03683899    : num  0.208 0.184 0.393 0.281 0.311 ...
##  $ cg03695421    : num  0.209 0.194 0.268 0.218 0.28 ...
##  $ cg03706273    : num  0.13 0.0985 0.0402 0.1015 0.0779 ...
```

We created our own codebook with a description per column. The details for the description where present on kaggle website for the dataset, but not in a codebook format.

```
code_book <- read.table(file = "archive/code_book.txt", sep = ";", header = T)
pander::pander(code_book)
```

column	description	type
GSM	The GSM with which the full sample data can be located on NCBI	factor
Smoking.status	Smoking status - never = never smoked current = currently smoking	factor

column	description	type
Gender	Patient Gender	factor
Age	Patient Age	numeric
cg00050873	Methylation in Current Island	numeric
cg00212031	Methylation in Current Island	numeric
cg00213748	Methylation in Current Island	numeric
cg00214611	Methylation in Current Island	numeric
cg00455876	Methylation in Current Island	numeric
cg01707559	Methylation in Current Island	numeric
cg02004872	Methylation in Current Island	numeric
cg02011394	Methylation in Current Island	numeric
cg02050847	Methylation in Current Island	numeric
cg02233190	Methylation in Current Island	numeric
cg02494853	Methylation in Current Island	numeric
cg02839557	Methylation in Current Island	numeric
cg02842889	Methylation in Current Island	numeric
cg03052502	Methylation in Current Island	numeric
cg03155755	Methylation in Current Island	numeric
cg03244189	Methylation in Current Island	numeric
cg03443143	Methylation in Current Island	numeric
cg03155755	Methylation in Current Island	numeric
cg03244189	Methylation in Current Island	numeric
cg03443143	Methylation in Current Island	numeric
cg03683899	Methylation in Current Island	numeric
cg03695421	Methylation in Current Island	numeric
cg03706273	Methylation in Current Island	numeric

When we look at the summary of the patient data we see in the column of the methylation data that there are 62 missing values. these are the same rows where the gender column has a capital F for female and M for male. these rows will be deleted because it gives us no information for the CpG sites.

```
#delete missing values
library(tidyr)
patient_data <- patient_data %>% drop_na()
pander::pander(apply(patient_data, 2, function(x) any(is.na(x))), caption = "Table to show per column wh
```

Table 8: Table continues below

GSM	Smoking.Status	Gender	Age	cg00050873	cg00212031	cg00213748
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table 9: Table continues below

cg00214611	cg00455876	cg01707559	cg02004872	cg02011394	cg02050847
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table 10: Table continues below

cg02233190	cg02494853	cg02839557	cg02842889	cg03052502	cg03155755
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

cg03244189	cg03443143	cg03683899	cg03695421	cg03706273
FALSE	FALSE	FALSE	FALSE	FALSE

```
# check if all column have no missing data
```

check if there is no missing data left in the other columns.

the Gender column contains abbreviations fore male and female. we changed this for the full name for a better readability.

```
str(patient_data)
```

```
## 'data.frame':    621 obs. of  24 variables:
##  $ GSM           : chr  "GSM1051525" "GSM1051526" "GSM1051527" "GSM1051528" ...
##  $ Smoking.Status: chr  "current" "current" "current" "current" ...
##  $ Gender        : chr  " f" " f" " f" " f" ...
##  $ Age           : int  67 49 53 62 33 59 66 51 55 37 ...
##  $ cg00050873    : num  0.608 0.345 0.321 0.277 0.414 ...
##  $ cg00212031    : num  0.423 0.569 0.361 0.304 0.131 ...
##  $ cg00213748    : num  0.372 0.501 0.353 0.475 0.368 ...
##  $ cg00214611    : num  0.622 0.499 0.374 0.486 0.761 ...
##  $ cg00455876    : num  0.291 0.375 0.231 0.295 0.236 ...
##  $ cg01707559    : num  0.267 0.19 0.315 0.296 0.251 ...
##  $ cg02004872    : num  0.179 0.156 0.106 0.111 0.169 ...
##  $ cg02011394    : num  0.48 0.418 0.615 0.301 0.393 ...
##  $ cg02050847    : num  0.328 0.346 0.238 0.305 0.306 ...
##  $ cg02233190    : num  0.241 0.175 0.246 0.177 0.302 ...
##  $ cg02494853    : num  0.0671 0.0469 0.0382 0.0267 0.037 ...
##  $ cg02839557    : num  0.24699 0.23674 0.24461 0.00164 0.33432 ...
##  $ cg02842889    : num  0.469 0.307 0.358 0.446 0.395 ...
##  $ cg03052502    : num  0.4 0.377 0.305 0.271 0.327 ...
##  $ cg03155755    : num  0.415 0.397 0.521 0.434 0.43 ...
##  $ cg03244189    : num  0.221 0.217 0.185 0.165 0.181 ...
##  $ cg03443143    : num  0.476 0.544 0.537 0.508 0.405 ...
##  $ cg03683899    : num  0.208 0.184 0.393 0.281 0.311 ...
##  $ cg03695421    : num  0.209 0.194 0.268 0.218 0.28 ...
##  $ cg03706273    : num  0.13 0.0985 0.0402 0.1015 0.0779 ...
```

```
library(dplyr)
```

```
##
## 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
```

```
library(dplyr)

patient_data <- patient_data %>% mutate(Gender=recode(Gender,
                          " f"="female",
                          " m"="male"))




#delete row GSM because this is a unique row identifier per patient and is no use for datamining

patient_data <- patient_data[-1]
head(patient_data[,1:4])
```

```
##   Smoking.Status Gender Age cg00050873
## 1        current female  67  0.6075634
## 2        current female  49  0.3450542
## 3        current female  53  0.3213497
## 4        current female  62  0.2772675
## 5          never female  33  0.4135991
## 6        current female  59  0.6228599
```

we also deleted the GMS column because is hold unique row identifiers. Fitting a tree with unique row identifiers in a dataset will split every single row in one node, which will give you a high predictive value. > This will cause overfitting.

Now that we took a general look at the data lets make some visualization to get a more in depth overview

```
library(ggplot2)



ggplot(patient_data, aes(x=Gender, y=Age, fill=Gender)) +
  geom_boxplot( )  + scale_fill_manual(values=c("pink", "royalblue")) +
  ggtitle("distribution of patient ages")+
  ylab("age in years")
```

if we compare the ages of patients like we did in this figure, we can see that the distribution of ages for male and female very alike

```
ggplot(data=patient_data, aes(Age)) +
  geom_histogram(fill='pink', color="black", alpha=0.3) +
  ggtitle("distribution of patient ages")+
  ylab("number of patients") + xlab("age in years")
```

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

Then the same age distribution but without seperation of the gender. We notice that the most patients are >45

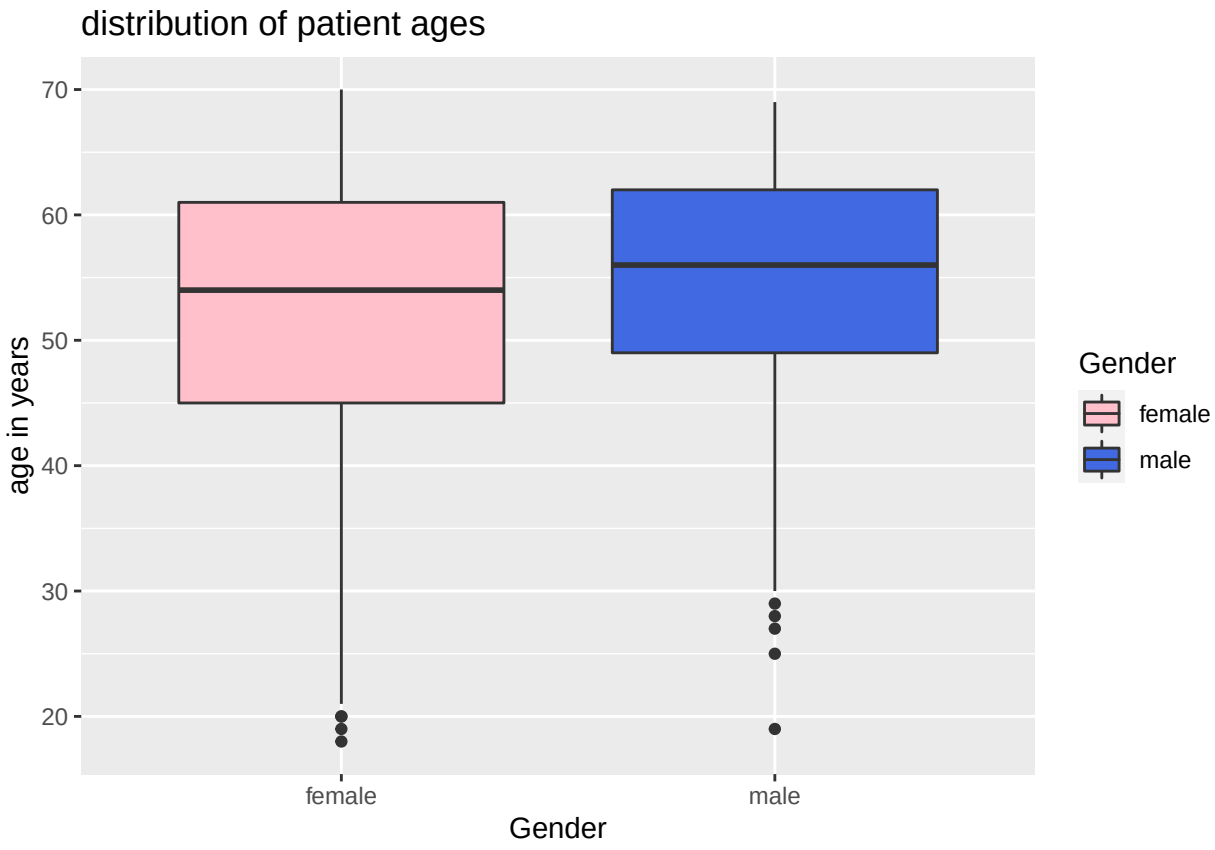
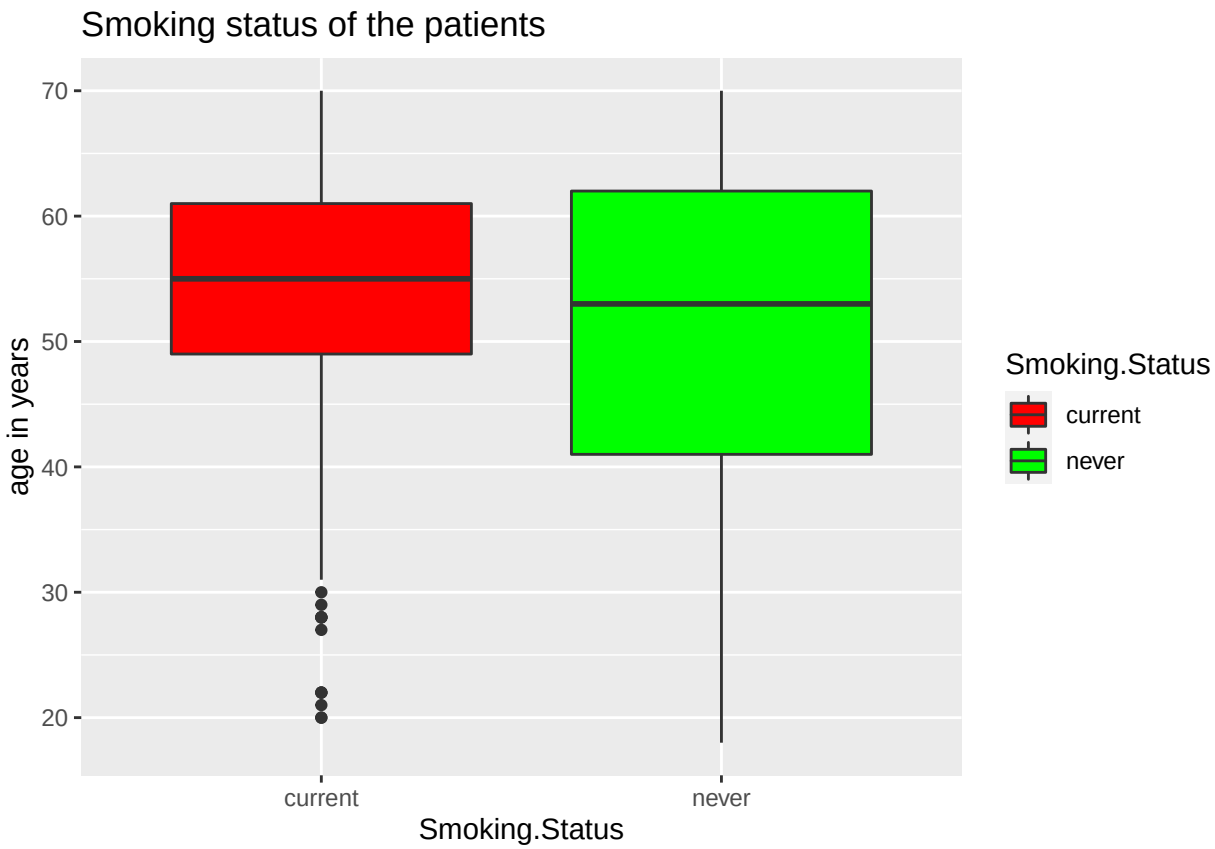


Figure 1: Title: age distribution of the patients compared to male and female

Figure 2: Title: age distribution of the patients without seperation of the gender

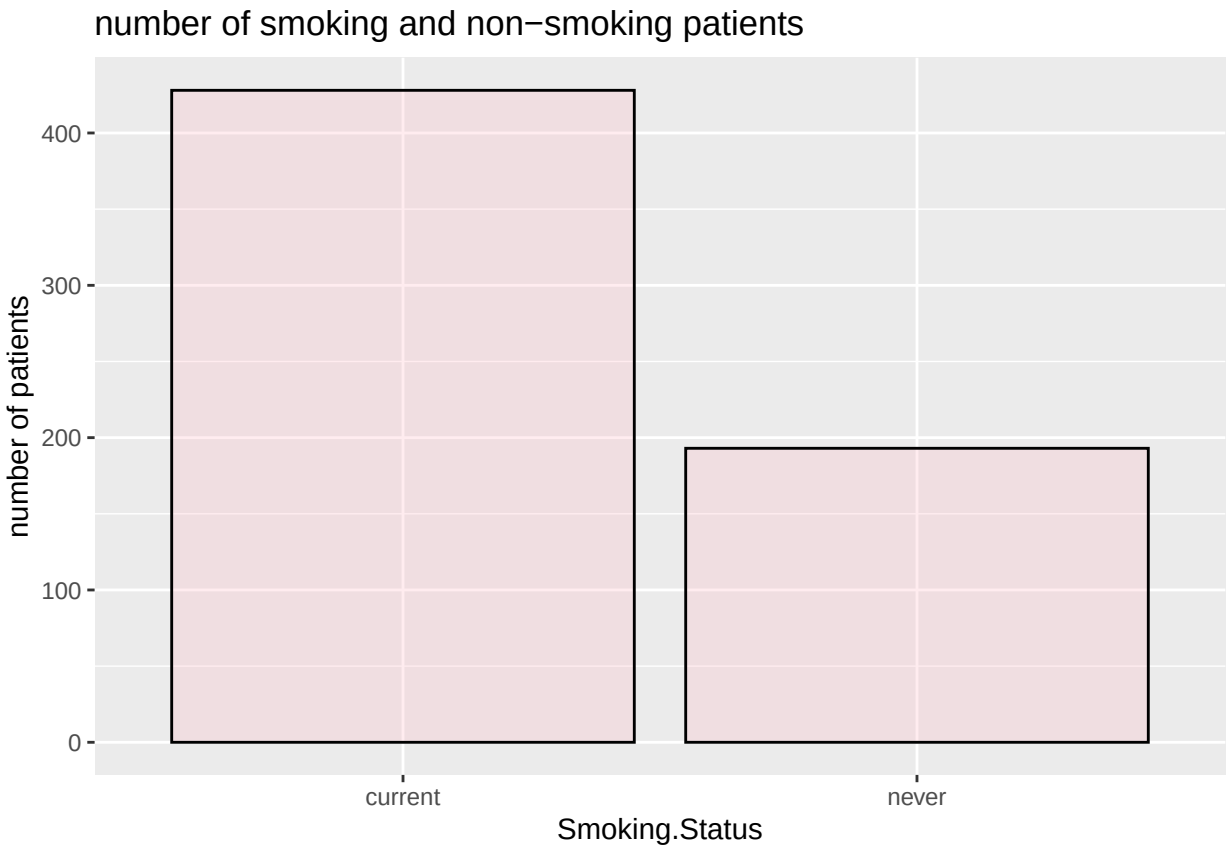
```
ggplot(patient_data, aes(x=Smoking.Status, y=Age, fill=Smoking.Status)) +
  geom_boxplot( )  + scale_fill_manual(values=c("red", "green")) +
  ggtitle("Smoking status of the patients")+
  ylab("age in years")
```



the younger patients in the dataset are non smoking.

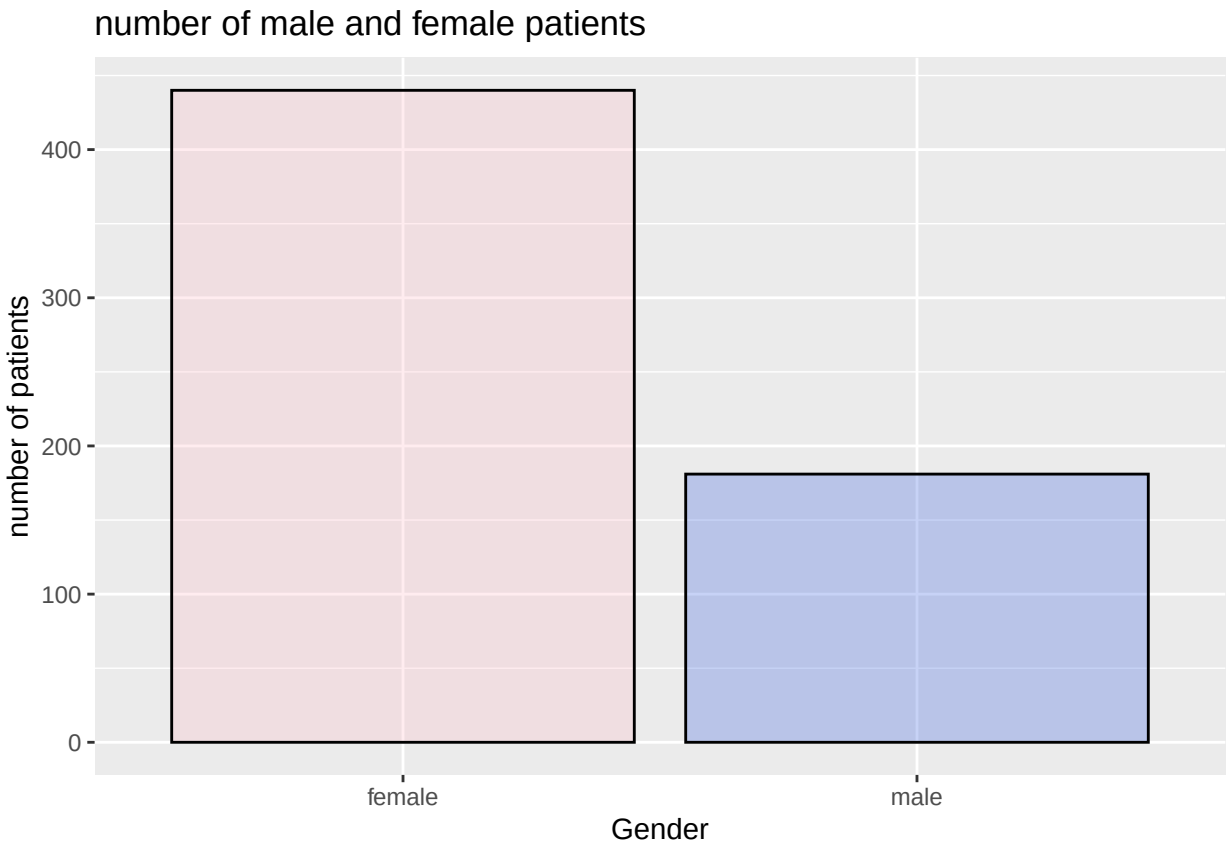
To get a quick overview of the most useful graphs with this data we use a ggpairs plot.

```
ggplot(data=patient_data, aes(Smoking.Status) ) +
  geom_bar(fill='pink', color="black", alpha=0.3) +
  ggtitle("number of smoking and non-smoking patients ")+
  ylab("number of patients")
```



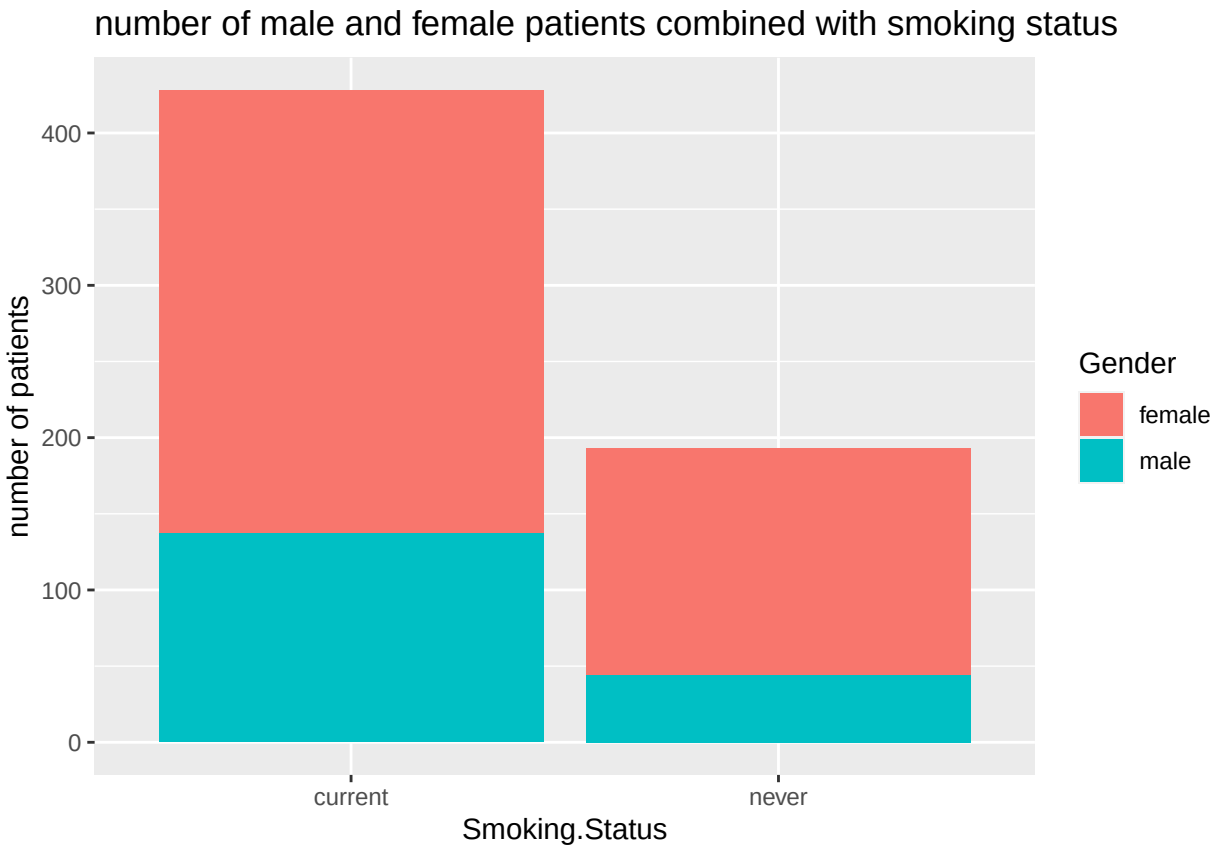
conclusion: high number of patients are smoking. We will not remove data to get an even distribution of smoking and non smoking, but we do need to keep this in mind when using datamining.

```
ggplot(data=patient_data, aes(Gender) ) +
  geom_bar(fill=c('pink',"royalblue"), color="black", alpha=0.3) +
  ggtitle("number of male and female patients ")+
  ylab("number of patients")
```



majority of the patients are female.

```
ggplot(data=patient_data, aes(Smoking.Status) ) + ggtitle("number of male and female patients combined w
  geom_bar(aes(fill=Gender)) +
  ylab("number of patients")
```



If we look a the two figures above is almost looks like the they are the same. and that all females are smoking and all males non smoking. So we made another plot to compare the both values in one.

Then we wanted to explore if there can be seen differences in the CpG values of males and females.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
long_data <- pivot_longer(data = patient_data, cols = 4:23, names_to = "body_part", values_to = "size")

long_data %>%  ggplot(aes(x = size,  colour = Gender)) +
    geom_density(show.legend = TRUE) +
  ggtitle("CpG values of males and females ") +
   facet_wrap(~body_part, ncol = 7) +    scale_color_manual(values=c("deeppink3", "royalblue") )
```

As you can see the figure, the male and female CpG values are very different from eachother. now we want to see if we get the same result for smoking and non smoking patients. which would mean that smoking changes your CpG values.

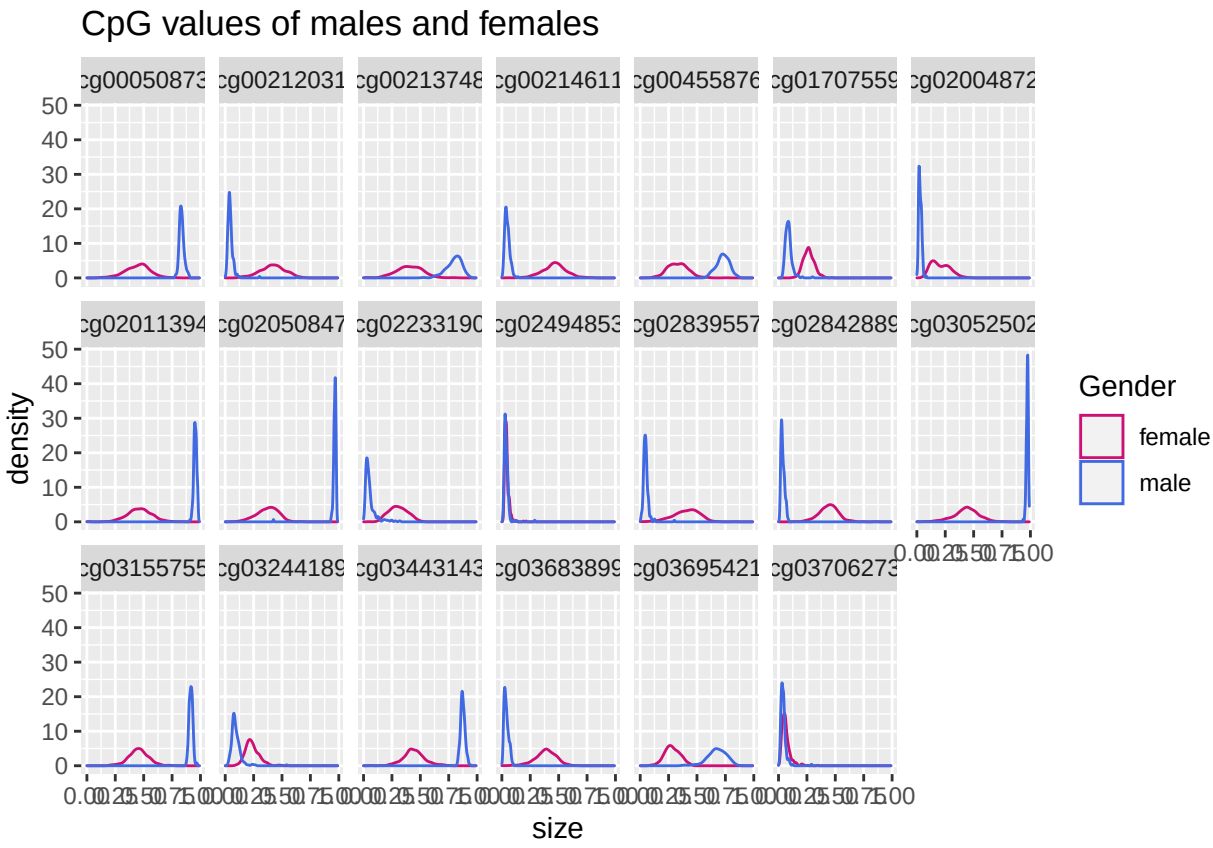


Figure 3: Title: CpG values in percentages comparison of males and females

```
long_data <- pivot_longer(data = patient_data, cols = 4:23, names_to = "body_part", values_to = "size")

long_data %>%  ggplot(aes(x = size,  colour = Smoking.Status)) +
    geom_density(show.legend = TRUE) +
  ggtitle("CpG values of smoking and non smoking patients ") +
   facet_wrap(~body_part, ncol = 5) +    scale_color_manual(values=c("red", "green"))
```

Figure 4: Title: CpG values in percentages comparison of smoking and non smoking patients

de dubbele curves zijn voor man en vrouw

annova test

standaard deviatie T-test per waarde om te zien of de lijen tussen roker en niet roker significant vershillen van elkaar of hetzelfde voor significiantie tussen man en vrouw