"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

```
options(width = 60)
matrix(runif(100), ncol = 20)
```

```
##           [,1]      [,2]       [,3]       [,4]      [,5]
## [1,] 0.7088905 0.6978659 0.18569808 0.12855795 0.2968125
## [2,] 0.3446528 0.4985605 0.72546016 0.70116697 0.1227183
## [3,] 0.9157977 0.1650005 0.46382856 0.15496871 0.4424312
## [4,] 0.4968653 0.9809454 0.88868768 0.11870698 0.8823486
## [5,] 0.8247535 0.8792306 0.01350082 0.07116472 0.1341742
##           [,6]       [,7]      [,8]       [,9]      [,10]
## [1,] 0.1995364 0.92331788 0.1370554 0.26198584 0.73879280
## [2,] 0.2727645 0.09120707 0.3612507 0.62918822 0.77207859
## [3,] 0.7654252 0.85667245 0.7795875 0.61712876 0.29570040
## [4,] 0.1337185 0.10380900 0.7875463 0.03969813 0.31406941
## [5,] 0.1371320 0.98210645 0.9822878 0.48512966 0.08335189
##           [,11]      [,12]     [,13]     [,14]     [,15]
## [1,] 0.81617855 0.91690077 0.5397503 0.4375926 0.9073288
## [2,] 0.32414595 0.83581065 0.5402197 0.7892517 0.2741032
## [3,] 0.06553951 0.08264142 0.2574242 0.2479113 0.2672167
## [4,] 0.05824456 0.13875933 0.1438382 0.3486492 0.5715743
## [5,] 0.11681374 0.37003380 0.8999927 0.6137524 0.2668780
##           [,16]     [,17]     [,18]      [,19]      [,20]
## [1,] 0.4411036 0.7512113 0.2717213 0.98070216 0.29832548
## [2,] 0.5804884 0.2546379 0.9889253 0.37304720 0.15516150
## [3,] 0.6246753 0.2293834 0.2460540 0.36186064 0.06664155
## [4,] 0.0444172 0.1266400 0.4758320 0.04435819 0.75724570
## [5,] 0.3769920 0.8854032 0.6540036 0.50820726 0.81181247
```

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

```
##          GSM Smoking.Status Gender Age cg00050873
## 1 GSM1051525        current      f  67  0.6075634
## 2 GSM1051526        current      f  49  0.3450542
## 3 GSM1051527        current      f  53  0.3213497
## 4 GSM1051528        current      f  62  0.2772675
## 5 GSM1051529          never      f  33  0.4135991
## 6 GSM1051530        current      f  59  0.6228599
##   cg00212031 cg00213748 cg00214611 cg00455876 cg01707559
## 1  0.4228427  0.3724549  0.6215619  0.2907773  0.2671431
## 2  0.5686615  0.5005995  0.4986067  0.3745909  0.1902743
## 3  0.3609091  0.3527315  0.3738240  0.2306740  0.3147052
## 4  0.3044371  0.4752352  0.4862581  0.2951815  0.2957931
## 5  0.1312511  0.3675446  0.7611667  0.2357703  0.2505265
## 6  0.5016849  0.2632270  0.4157459  0.4751891  0.2539041
##   cg02004872 cg02011394 cg02050847 cg02233190 cg02494853
## 1  0.1791439  0.4802517  0.3276078  0.2411204 0.06706958
## 2  0.1559775  0.4180809  0.3464627  0.1754907 0.04693889
## 3  0.1057448  0.6151030  0.2375392  0.2464092 0.03823712
## 4  0.1112862  0.3010196  0.3045353  0.1770279 0.02671625
## 5  0.1691084  0.3929746  0.3062257  0.3017014 0.03701636
## 6  0.2607587  0.5097921  0.4052457  0.3852716 0.02583463
##    cg02839557 cg02842889 cg03052502 cg03155755 cg03244189
## 1 0.246993368  0.4692396  0.4002466  0.4150313  0.2214331
## 2 0.236742313  0.3074666  0.3770313  0.3973715  0.2171221
## 3 0.244611725  0.3577526  0.3050442  0.5212775  0.1850495
```

```
## 4 0.001641439  0.4457390  0.2714746  0.4344920  0.1654187
## 5 0.334319727  0.3950396  0.3265530  0.4300966  0.1811352
## 6 0.309210202  0.3218573  0.5333670  0.5715522  0.2109749
##   cg03443143 cg03683899 cg03695421 cg03706273
## 1  0.4758258  0.2077242  0.2091974 0.12998255
## 2  0.5444690  0.1844462  0.1937732 0.09853265
## 3  0.5370600  0.3931231  0.2680030 0.04024808
## 4  0.5079167  0.2812089  0.2178572 0.10151626
## 5  0.4054791  0.3107944  0.2800708 0.07785712
## 6  0.3778239  0.4693609  0.3433317 0.04577912
```

```
#View(patient_data)

knitr::kable(summary(patient_data), caption = " CpG values dataset ")
```

GSM	Smoking.Status	Gender	Age	cg00050873	cg00212031	cg00213748	
GSM2219538: 2	current:490	f:440	Min. :18.00	Min. :0.1186	Min. :0.00695	Min. :0.0000	M
GSM2219539: 2	never :193	F: 13	1st Qu.:47.00	1st Qu.:0.4131	1st Qu.:0.06317	1st Qu.:0.3635	1s
GSM2219540: 2	NA	m:181	Median :56.00	Median :0.5052	Median :0.36554	Median :0.4713	Me
GSM2219541: 2	NA	M: 49	Mean :53.82	Mean :0.5600	Mean :0.30960	Mean :0.5191	M
GSM2219542: 2	NA	NA	3rd Qu.:62.00	3rd Qu.:0.8144	3rd Qu.:0.45981	3rd Qu.:0.7278	3r
GSM2219543: 2	NA	NA	Max. :80.00	Max. :0.8989	Max. :0.70999	Max. :0.9236	M
(Other) :671	NA	NA	NA	NA's :62	NA's :62	NA's :62	

```
str(patient_data)
```

```
## 'data.frame':    683 obs. of  24 variables:
##  $ GSM           : Factor w/ 671 levels "GSM1051525","GSM1051526",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Smoking.Status: Factor w/ 2 levels "current","never": 1 1 1 1 2 1 2 1 1 2 ...
##  $ Gender        : Factor w/ 4 levels " f"," F"," m",..: 1 1 1 1 1 1 1 1 3 3 ...
##  $ 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)
kable(code_book, caption = "A codebook for the data ")
```

Table 2: A codebook for the data

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
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()

# check if all column have no missing data
kable(apply(patient_data, 2, function(x) any(is.na(x))), caption = "Table to show per column whether th
```

Table 3: Table to show per column whether there are missing values
FALSE= no missing values found / TRUE = missing values found

	x
GSM	FALSE
Smoking.Status	FALSE
Gender	FALSE
Age	FALSE
cg00050873	FALSE
cg00212031	FALSE
cg00213748	FALSE
cg00214611	FALSE
cg00455876	FALSE
cg01707559	FALSE
cg02004872	FALSE
cg02011394	FALSE
cg02050847	FALSE
cg02233190	FALSE
cg02494853	FALSE
cg02839557	FALSE
cg02842889	FALSE
cg03052502	FALSE
cg03155755	FALSE
cg03244189	FALSE
cg03443143	FALSE
cg03683899	FALSE
cg03695421	FALSE
cg03706273	FALSE

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           : Factor w/ 671 levels "GSM1051525","GSM1051526",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Smoking.Status: Factor w/ 2 levels "current","never": 1 1 1 1 2 1 2 1 1 2 ...
##  $ Gender        : Factor w/ 4 levels " f"," F"," m",..: 1 1 1 1 1 1 1 1 3 3 ...
##  $ 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
```

```
levels(patient_data$Gender) <- c("female", "female", "male", "male")




#head(patient_data)

# delete rij GSM aangezien deze uniek zijn en dit niet gaat werken met 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")
```

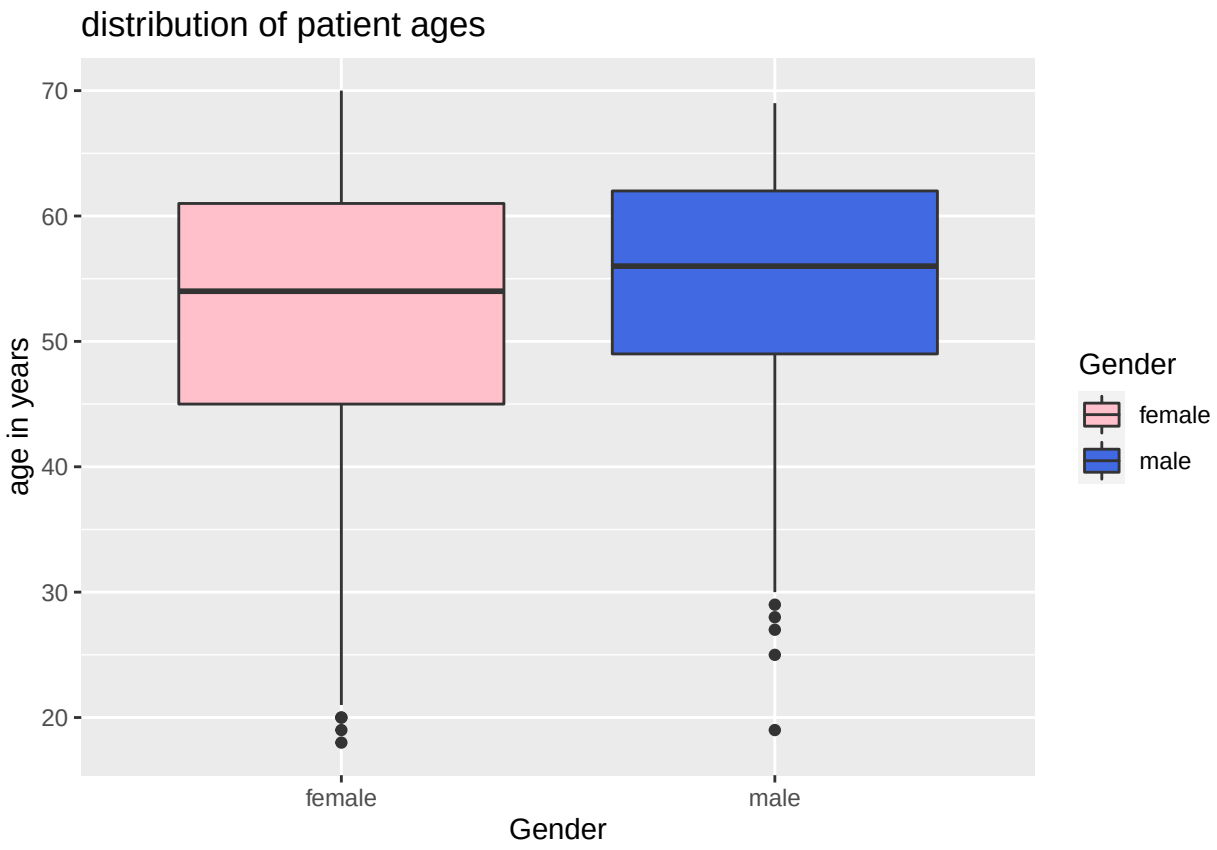


Figure 1: Title: age distribution of the patients compared to male and female

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

```
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")
```

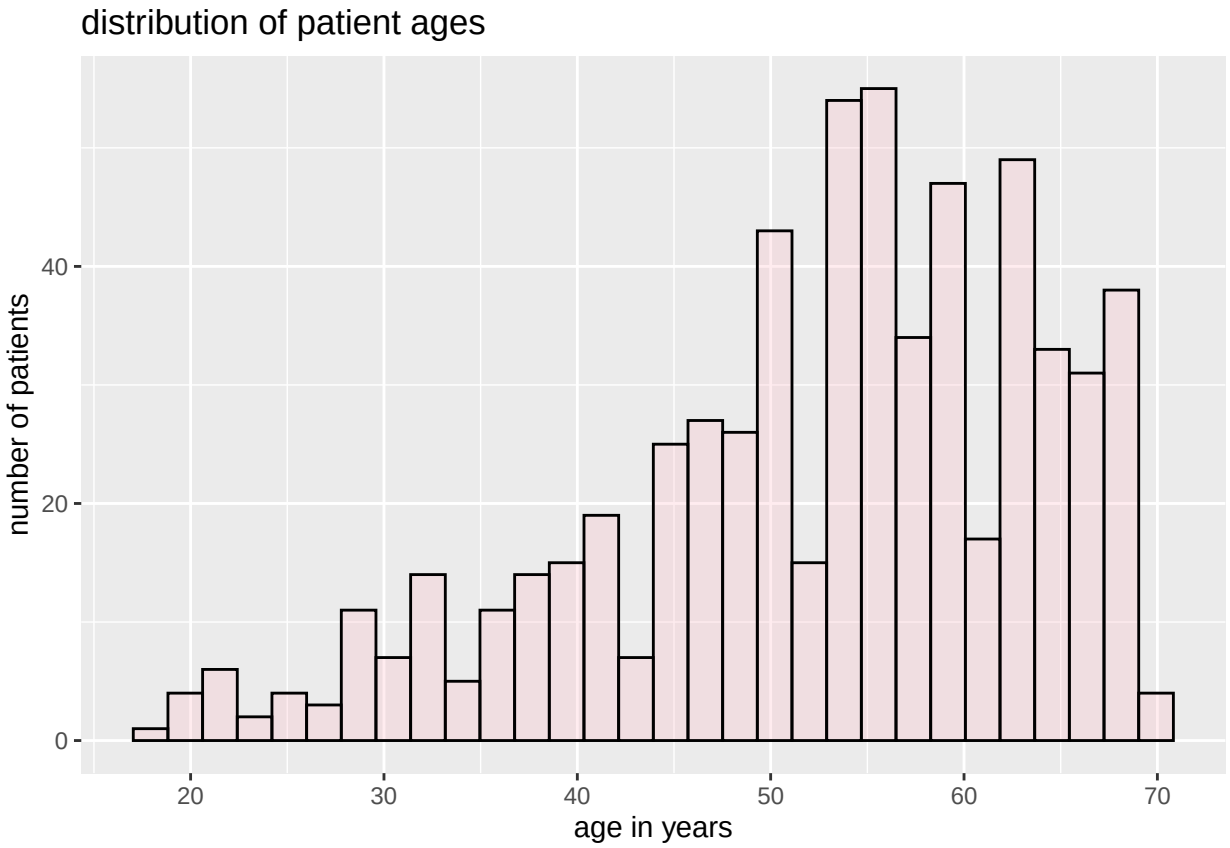
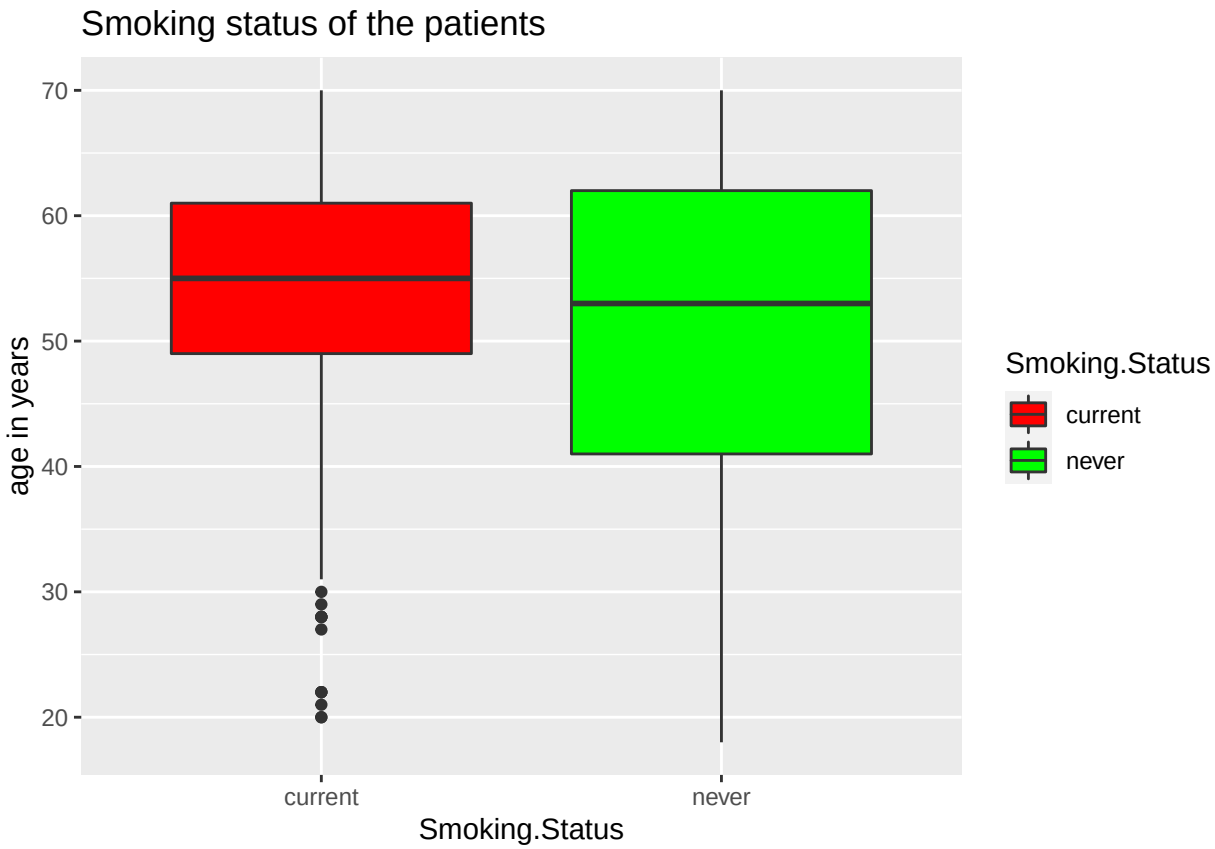


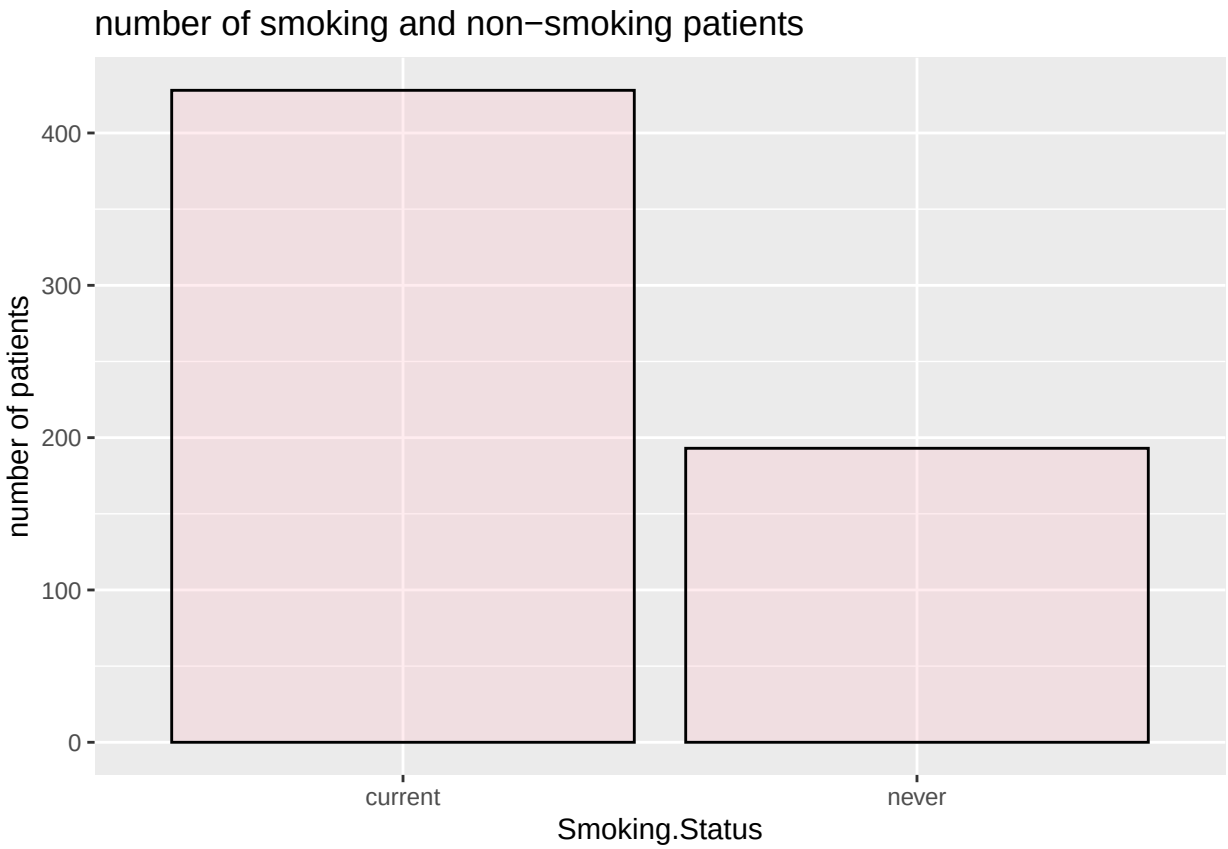
Figure 2: Title: age distribution of the patients without seperation of the gender



the younger patients in the dataset are non smoking.

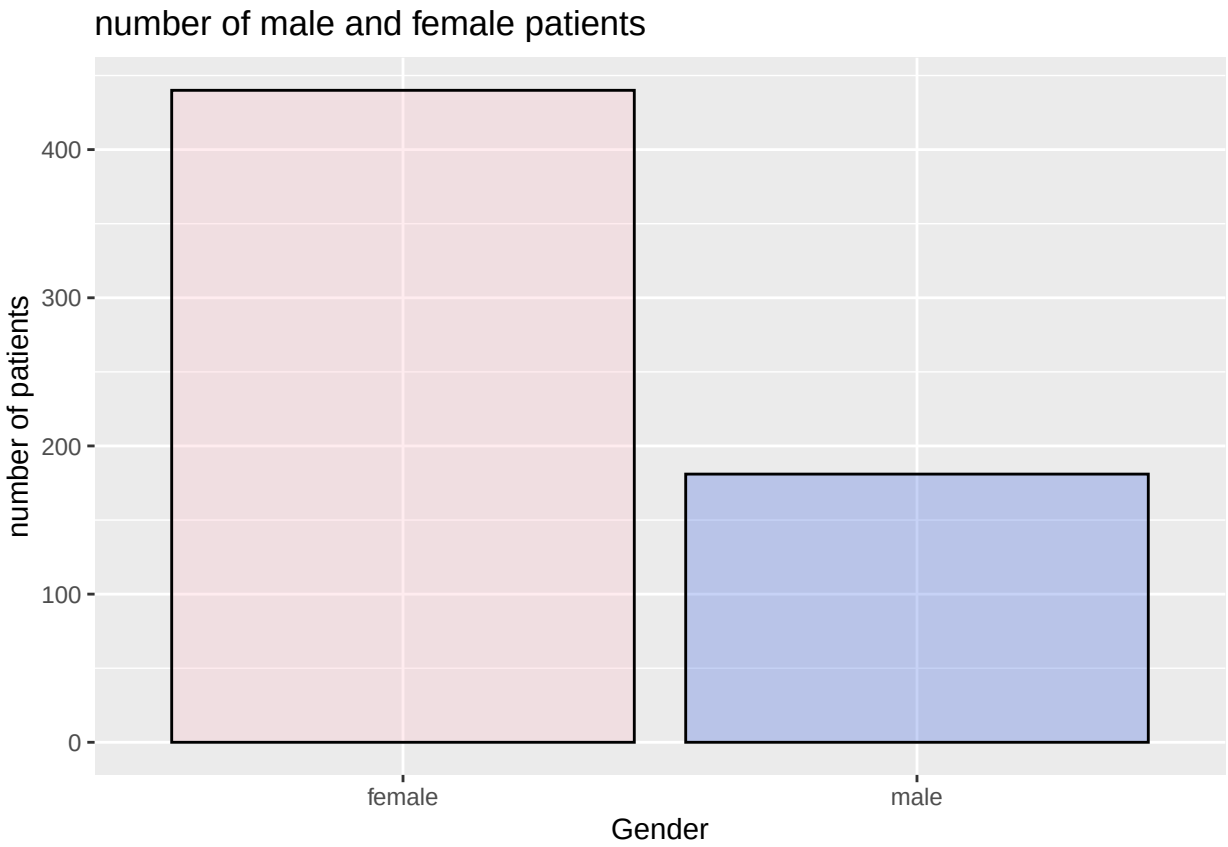
To get a quick overview of the most useful graphs with this data we use a ggpairs plot.

```
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")
```



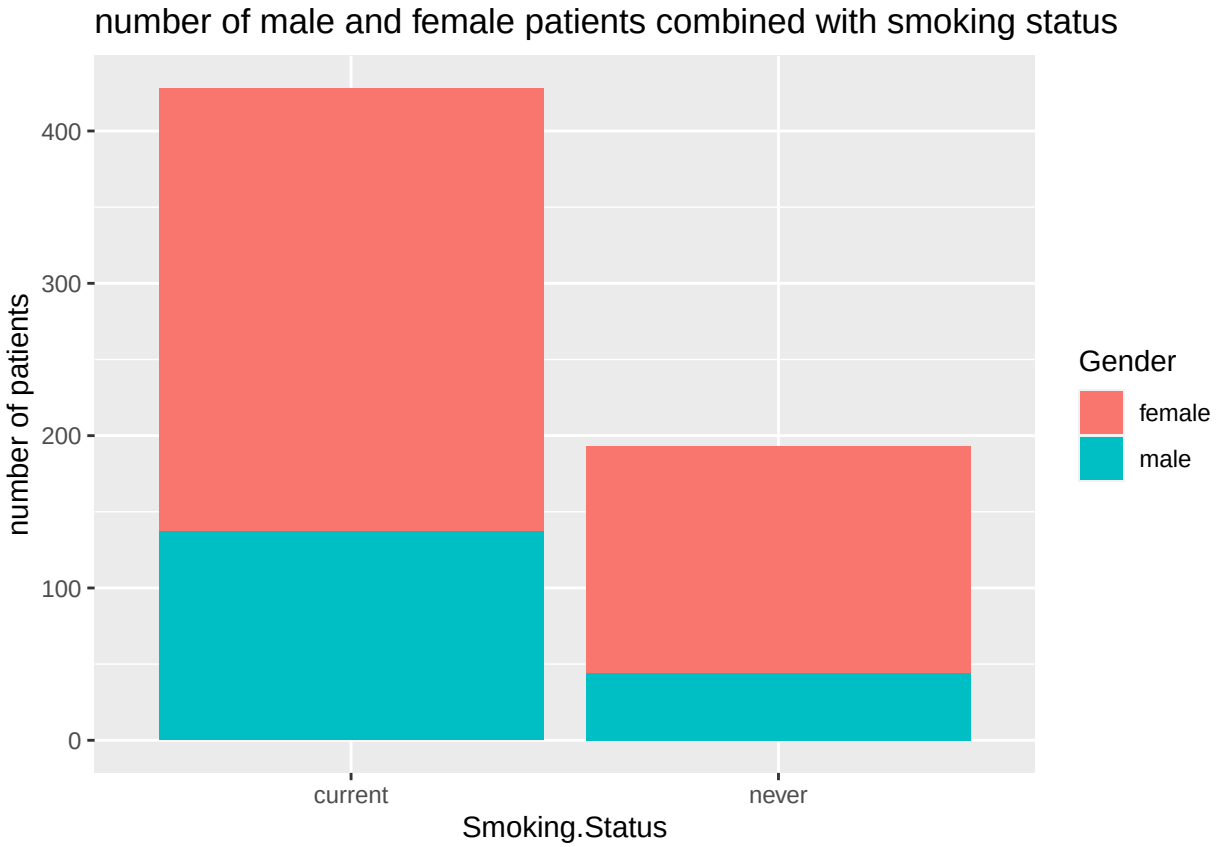
conclusion: high number of patients are smoking. We will not remove data to get an even distribution of smoking and non smoking, but we do need to keep this in mind when using datamining.

```
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 1.3.0 --

## v tibble  3.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.3

## -- 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.

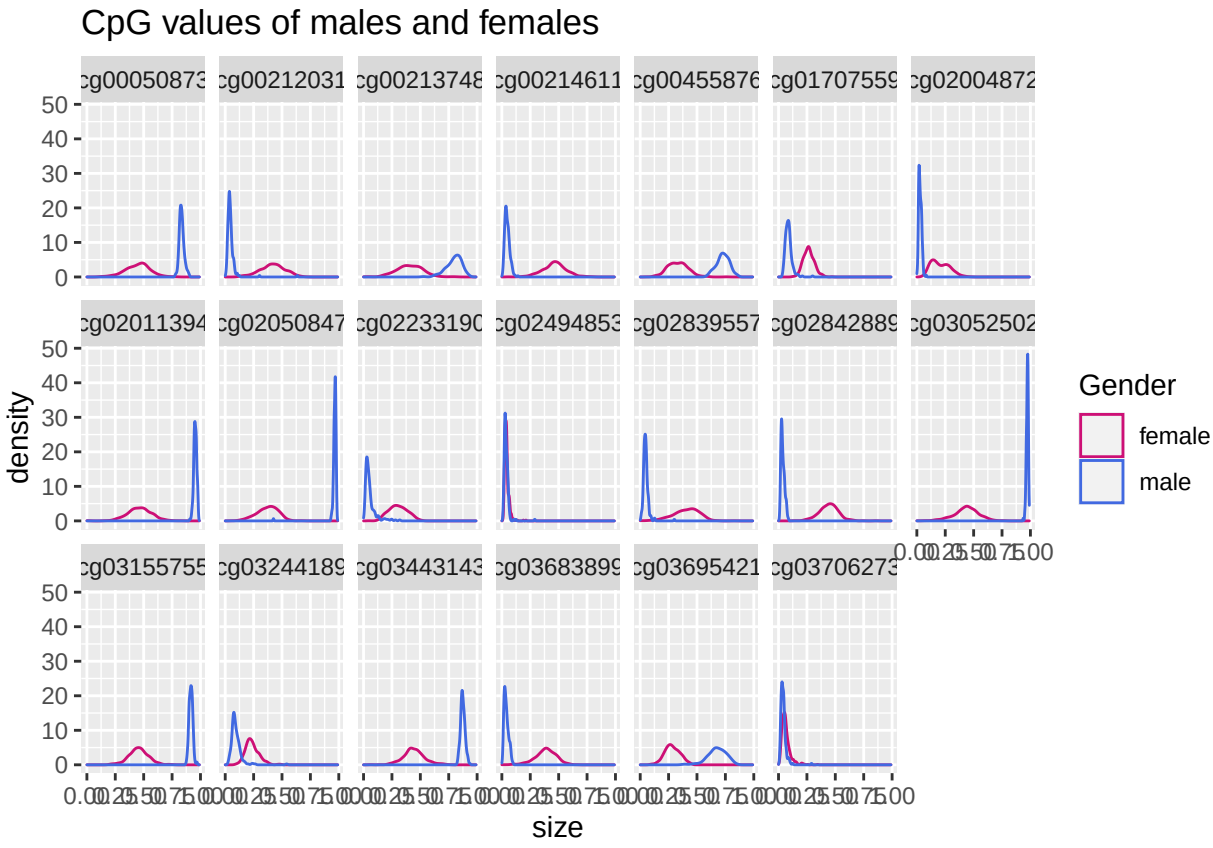


Figure 3: Title: CpG values in percentages comparison of males and females

```
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