

Preliminary Investigation

Ethan Allavarpu (UID: 405287603)

10/27/2020

```
sample <- read.csv("sample.csv", stringsAsFactors = TRUE)
# sample

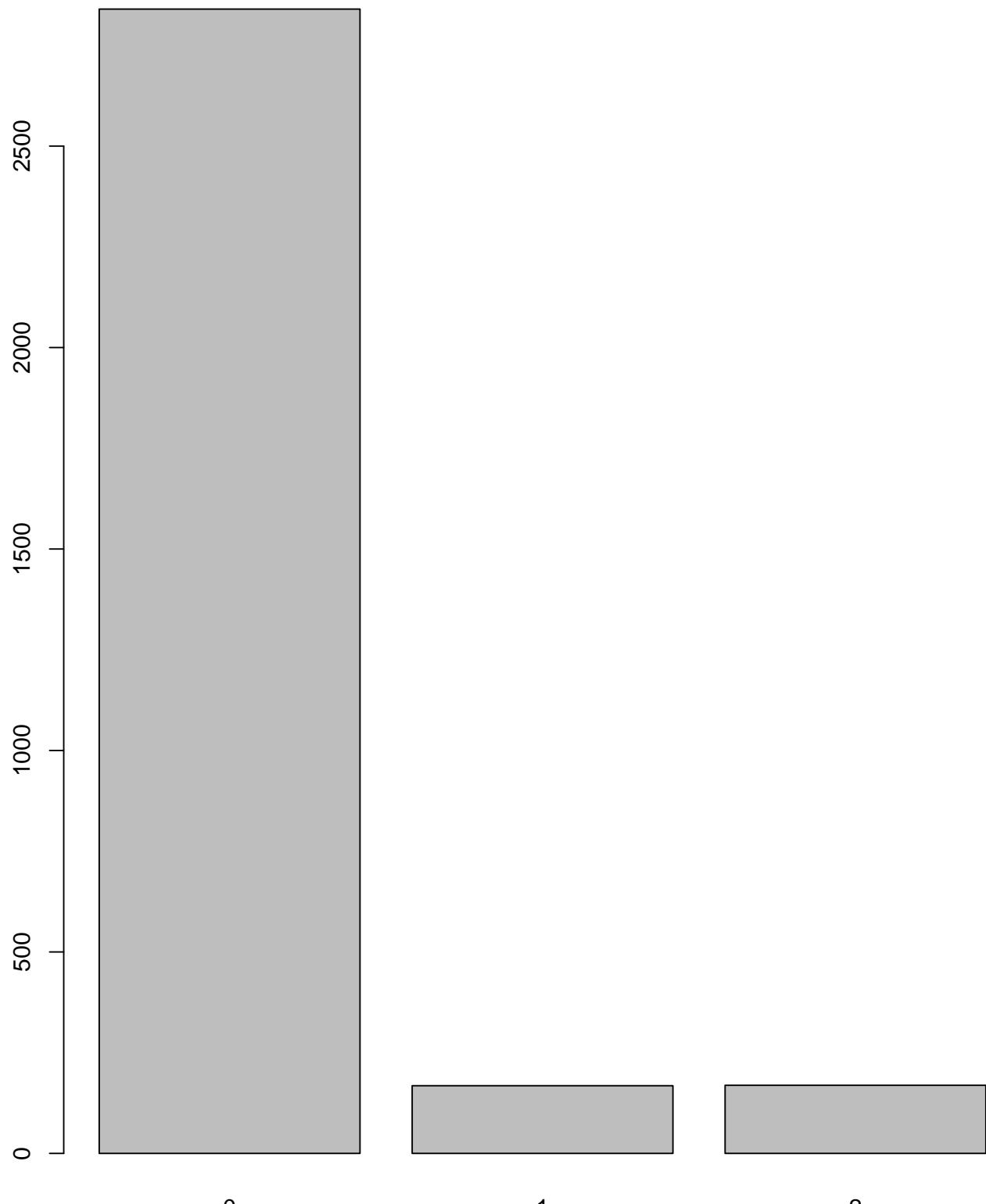
training <- read.csv("training.csv", stringsAsFactors = TRUE)
dim(training)

[1] 3177    99
names(training)[c(1, 99)]

[1] "id"      "class"
barplot(table(training$class))
table(training$class) / nrow(training)

0          1          2
0.89392509 0.05288008 0.05319484
any(is.na(training))

[1] FALSE
library(ggplot2)
```

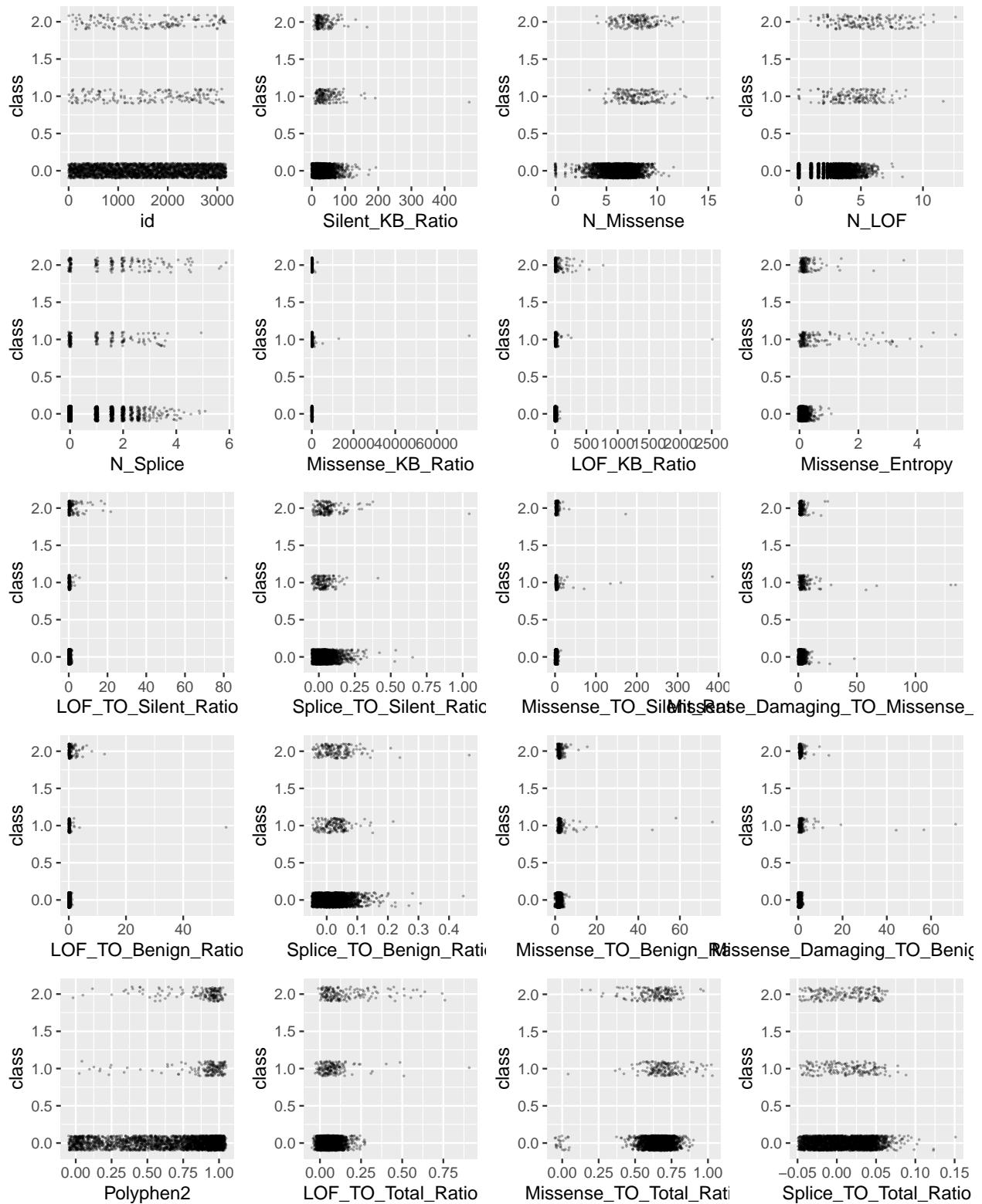


```
scatter <- function(var) {  
  ggplot(training, aes_string(var, "class")) +  
    geom_jitter(width = 0.05, height = 0.1, size = 0.1,  
                colour = rgb(0, 0, 0, alpha = 1 / 3))  
}
```

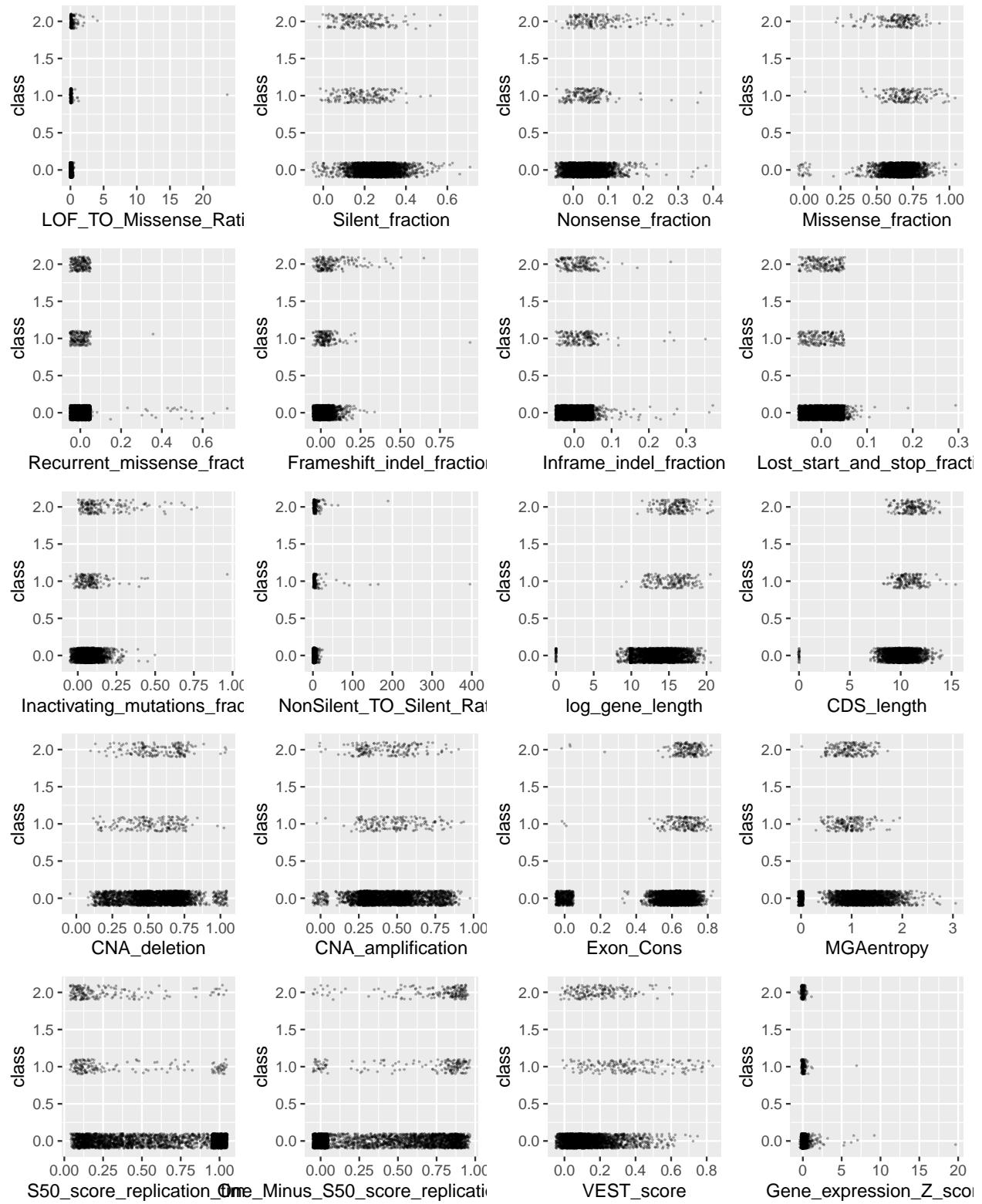
```

scat_plot <- lapply(names(training)[-99], scatter)
library(gridExtra)
grid.arrange(grobs = scat_plot[1:20], ncol = 4)

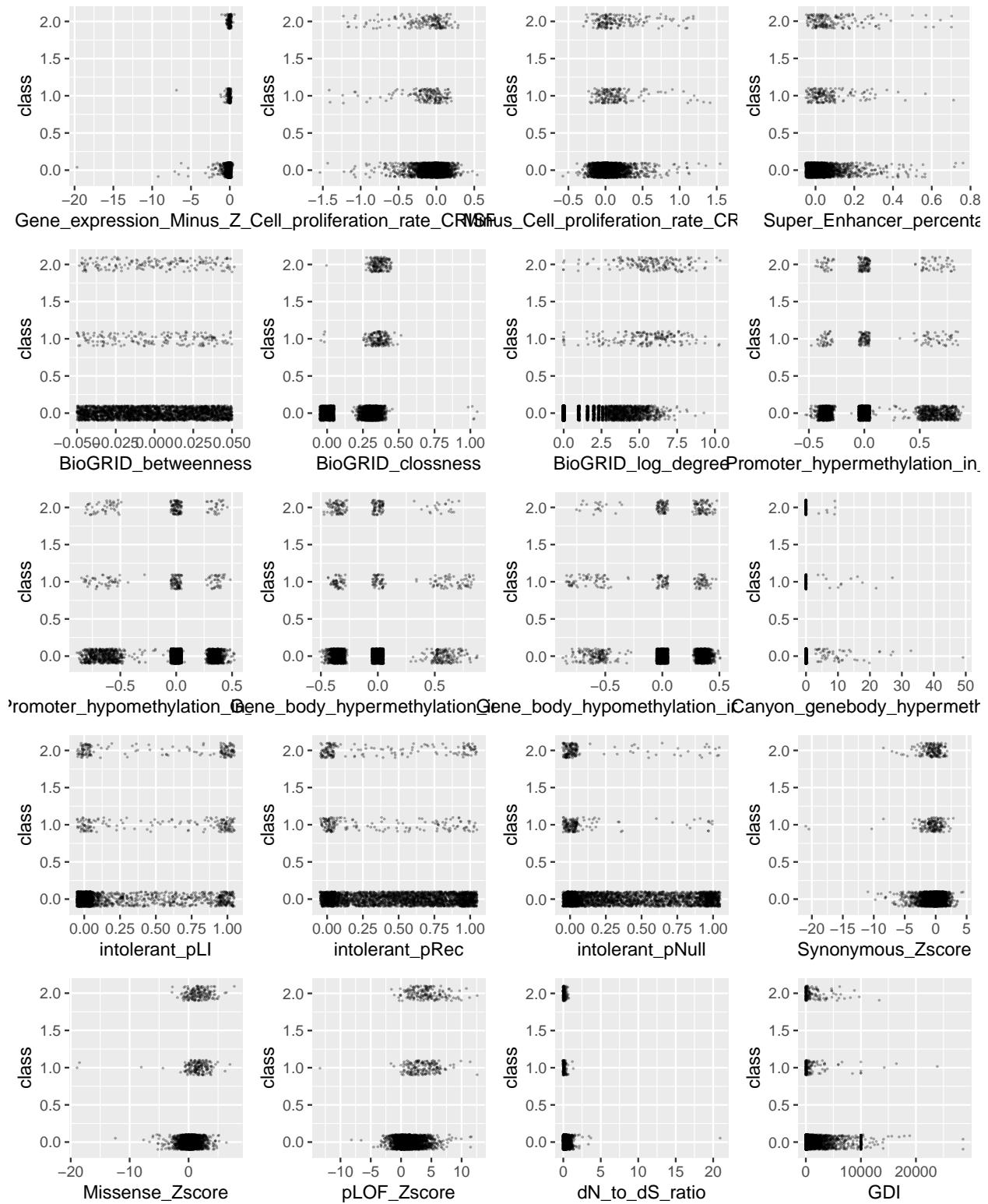
```



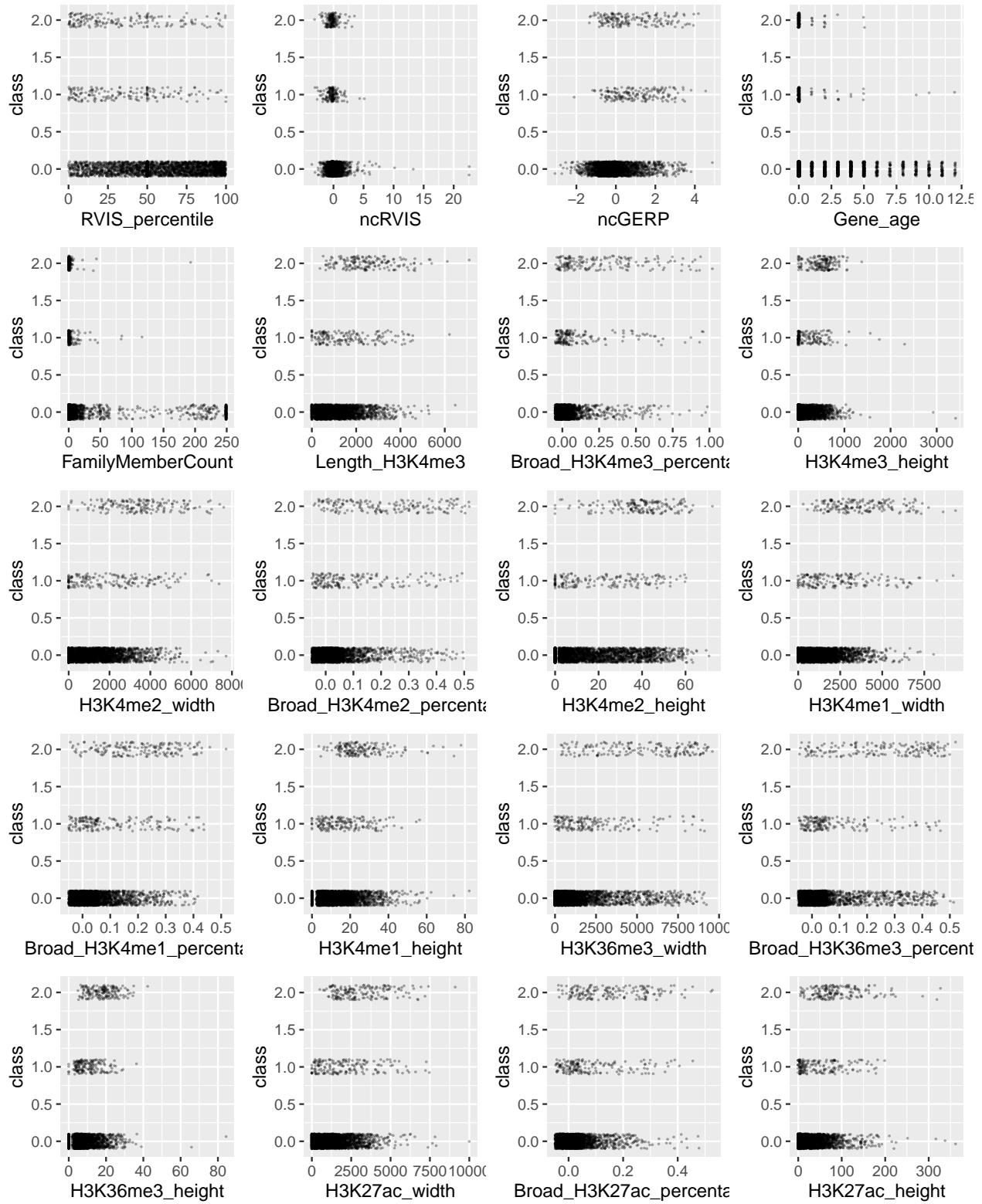
```
grid.arrange(grobs = scat_plot[21:40], ncol = 4)
```



```
grid.arrange(grobs = scat_plot[41:60], ncol = 4)
```



```
grid.arrange(grobs = scat_plot[61:80], ncol = 4)
```



```
grid.arrange(grobs = scat_plot[81:98], ncol = 4)
```

