Data_Analysis

Rienk Heins

3/3/2022

```
library(flowCore)
library(ggcyto)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ncdfFlow
```

```
## Loading required package: RcppArmadillo
```

```
## Loading required package: BH
```

```
## Loading required package: flowWorkspace
```

```
## As part of improvements to flowWorkspace, some behavior of
## GatingSet objects has changed. For details, please read the section
## titled "The cytoframe and cytoset classes" in the package vignette:
##
##   vignette("flowWorkspace-Introduction", "flowWorkspace")
```

```
library(ggpubr)
library(Biobase)
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following object is masked from 'package:flowCore':
##
##     normalize
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
library(ggplot2)
```

Loading the data

Dataset 1:

```
file_list1 <- list.files(path="../FACS_Data/PHBV_Micha1/")
dataset1 <- data.frame()
for(i in 1:length(file_list1)){
  file_string <- file_list1[i]
  file_string <- paste("../FACS_Data/PHBV_Micha1/", file_string, sep = "")
  fcs.data <- read.FCS(file_string)
  fcs.data.frame <- as.data.frame(exprs(fcs.data))
  dataset1 <- rbind(dataset1, fcs.data.frame)
}
head(dataset1)
```

```
##     FSC-A  SSC-A FL1-A FL2-A FL3-A FL4-A  FSC-H   SSC-H FL1-H FL2-H FL3-H FL4-H
## 1  183714  83123    92   109    94   249 178582  108792    69    64   166    37
## 2  135029  50422   125   219   284   467 120138   64549   185   145   382   189
## 3 1152415 963584   136    51   382   354 962586 1128253   180    43   246   145
## 4  471739 125265     4    27    87   238 458803  162676    59    74   220    28
## 5  432272 103531   127   113   259   463 397744  130475   105   112   185   417
## 6  158793  59836   185    99   363   290 145732   75252    97   107   345   260
##   Width Time
## 1    45  138
## 2    43  138
## 3    70  138
## 4    56  138
## 5    56  138
## 6    44  138
```

```
summary(dataset1)
```

```
##      FSC-A              SSC-A              FL1-A            FL2-A
```

```
##  Min.   :   71520   Min.   :       0   Min.   :     0   Min.   :     0
##  1st Qu.:  151637   1st Qu.:   53169   1st Qu.:    98   1st Qu.:    80
##  Median :  284776   Median :  105169   Median :   252   Median :   160
##  Mean   :  625561   Mean   :  488612   Mean   :  1317   Mean   :   695
##  3rd Qu.:  621818   3rd Qu.:  267424   3rd Qu.:   653   3rd Qu.:   363
##  Max.   :16777215   Max.   :16777215   Max.   :909767   Max.   :756424
##      FL3-A             FL4-A              FSC-H             SSC-H
##  Min.   :      0   Min.   :     0.0   Min.   :  80002   Min.   :       0
##  1st Qu.:    133   1st Qu.:   239.0   1st Qu.: 125199   1st Qu.:   64622
##  Median :    275   Median :   326.0   Median : 233018   Median :  122023
##  Mean   :   1103   Mean   :   443.1   Mean   : 429632   Mean   :  434213
##  3rd Qu.:    602   3rd Qu.:   450.0   3rd Qu.: 523223   3rd Qu.:  279105
##  Max.   :1395513   Max.   :851444.0   Max.   :8772799   Max.   :16777215
##      FL1-H             FL2-H              FL3-H             FL4-H
##  Min.   :     0.0   Min.   :     0.0   Min.   :     0.0   Min.   :     0
##  1st Qu.:   101.0   1st Qu.:    82.0   1st Qu.:   162.0   1st Qu.:   106
##  Median :   211.0   Median :   134.0   Median :   282.0   Median :   202
##  Mean   :   717.6   Mean   :   382.9   Mean   :   680.2   Mean   :   322
##  3rd Qu.:   466.0   3rd Qu.:   257.0   3rd Qu.:   530.0   3rd Qu.:   354
##  Max.   :192095.0   Max.   :100616.0   Max.   :434806.0   Max.   :668152
##      Width              Time
##  Min.   :  27.00   Min.   :  136
##  1st Qu.:  46.00   1st Qu.:  210
##  Median :  55.00   Median :  352
##  Mean   :  60.83   Mean   : 1702
##  3rd Qu.:  66.00   3rd Qu.:  515
##  Max.   :1001.00   Max.   :21344
```

Dataset 2:

```
file_list2 <- list.files(path="../FACS_Data/PHBV_Micha2/")
dataset2 <- data.frame()
for(i in 1:length(file_list2)){
  file_string <- file_list2[i]
  file_string <- paste("../FACS_Data/PHBV_Micha2/", file_string, sep = "")
  fcs.data <- read.FCS(file_string)
  fcs.data.frame <- as.data.frame(exprs(fcs.data))
  dataset2 <- rbind(dataset2, fcs.data.frame)
}
head(dataset2)
```

```
##    FSC-A  SSC-A FL1-A FL2-A FL3-A FL4-A  FSC-H  SSC-H FL1-H FL2-H FL3-H FL4-H
## 1  84384  17074   178   193   221   290  84405  23544   172   212   215   176
## 2 305008  28374   266   175   396   351 306267  38654   305   184   393   231
## 3 319879 103346   122   129   376   320 294995 134801   141   122   266   161
## 4 458032  65853    56   106    94   449 448352  82397    80   116   111   348
## 5 948341 143033  1353   961  2209  1330 866414 168838  1153   651  1937  1098
## 6 615074 100811   108    46   379    84 609837 129923    95    88   452   108
##   Width Time
## 1    33  137
## 2    51  137
## 3    53  137
## 4    56  137
```

```
## 5    66  137
## 6    59  137
```

```
summary(dataset2)
```

```
##      FSC-A              SSC-A              FL1-A               FL2-A
##  Min.   :   78317   Min.   :       0   Min.   :      0.0   Min.   :      0.0
##  1st Qu.:  142316   1st Qu.:   15417   1st Qu.:     62.0   1st Qu.:     56.0
##  Median :  169082   Median :   39270   Median :     94.0   Median :     93.0
##  Mean   :  361979   Mean   :  133875   Mean   :    368.7   Mean   :    250.3
##  3rd Qu.:  424243   3rd Qu.:   87175   3rd Qu.:    167.0   3rd Qu.:    151.0
##  Max.   :16777215   Max.   :16777215   Max.   :1932365.0   Max.   :1124103.0
##      FL3-A              FL4-A              FSC-H              SSC-H
##  Min.   :     0.0   Min.   :     0.0   Min.   :  80000   Min.   :       0
##  1st Qu.:   111.0   1st Qu.:   240.0   1st Qu.: 134528   1st Qu.:   20324
##  Median :   194.0   Median :   308.0   Median : 161208   Median :   50383
##  Mean   :   585.2   Mean   :   370.9   Mean   : 302485   Mean   :  137355
##  3rd Qu.:   353.0   3rd Qu.:   397.0   3rd Qu.: 381769   3rd Qu.:  107829
##  Max.   :761278.0   Max.   :125819.0   Max.   :7610997   Max.   :16573002
##      FL1-H              FL2-H              FL3-H              FL4-H
##  Min.   :     0.0   Min.   :     0.0   Min.   :     0.0   Min.   :    0.0
##  1st Qu.:    66.0   1st Qu.:    63.0   1st Qu.:   131.0   1st Qu.:   73.0
##  Median :    96.0   Median :    91.0   Median :   217.0   Median :  152.0
##  Mean   :   237.2   Mean   :   168.7   Mean   :   441.5   Mean   :  229.2
##  3rd Qu.:   159.0   3rd Qu.:   136.0   3rd Qu.:   369.0   3rd Qu.:  270.0
##  Max.   :512534.0   Max.   :302385.0   Max.   :257022.0   Max.   :69399.0
##      Width              Time
##  Min.   :  27.00   Min.   :   136
##  1st Qu.:  42.00   1st Qu.:   232
##  Median :  44.00   Median :   326
##  Mean   :  51.38   Mean   :  4043
##  3rd Qu.:  56.00   3rd Qu.:   483
##  Max.   :1001.00   Max.   :102145
```

Dataset 3:

```
file_list3 <- list.files(path="../FACS_Data/PHBV_Micha3/")
dataset3 <- data.frame()
for(i in 1:length(file_list3)){
  file_string <- file_list3[i]
  file_string <- paste("../FACS_Data/PHBV_Micha3/", file_string, sep = "")
  fcs.data <- read.FCS(file_string)
  fcs.data.frame <- as.data.frame(exprs(fcs.data))
  dataset3 <- rbind(dataset3, fcs.data.frame)
}
head(dataset3)
```

```
##    FSC-A  SSC-A FL1-A FL2-A FL3-A FL4-A  FSC-H   SSC-H FL1-H FL2-H FL3-H FL4-H
## 1 221141  73923   210   118   360   406 189843   92589   205    79   389   259
## 2 553662  61085    54    46   466   716 558818   81041    75    78   491   385
## 3 694700  68253   169   166  2170  4739 697233   90823   141   114  2439 11938
## 4 974360 847839  2452  1970  6453  6670 802617 1038551  1988  1633  6031  5426
## 5 239358  26945   125   120   382  1707 232815   36891   167    56   525  1415
```

```
## 6 426962  41192    52    93   451   832 419129   54988    53    90   482   858
##   Width Time
## 1    50  366
## 2    57  366
## 3    59  366
## 4    67  366
## 5    49  366
## 6    55  367
```

```
summary(dataset3)
```

```
##      FSC-A             SSC-A             FL1-A             FL2-A
##  Min.   :   71942   Min.   :       0   Min.   :      0   Min.   :      0
##  1st Qu.:  156580   1st Qu.:   39004   1st Qu.:     80   1st Qu.:    100
##  Median :  295879   Median :   77507   Median :    159   Median :    190
##  Mean   :  540272   Mean   :  343454   Mean   :   1222   Mean   :   1207
##  3rd Qu.:  587482   3rd Qu.:  203239   3rd Qu.:    406   3rd Qu.:    429
##  Max.   :16777215   Max.   :16777215   Max.   :4201918   Max.   :2544087
##      FL3-A             FL4-A             FSC-H             SSC-H
##  Min.   :      0   Min.   :     0   Min.   :  80000   Min.   :       0
##  1st Qu.:    441   1st Qu.:   314   1st Qu.: 128274   1st Qu.:   48366
##  Median :    973   Median :   533   Median : 243145   Median :   90748
##  Mean   :   5390   Mean   :  1520   Mean   : 390563   Mean   :  315598
##  3rd Qu.:   2316   3rd Qu.:  1222   3rd Qu.: 509813   3rd Qu.:  207752
##  Max.   :9497110   Max.   :643928   Max.   :8683647   Max.   :16777215
##      FL1-H             FL2-H              FL3-H             FL4-H
##  Min.   :     0   Min.   :     0.0   Min.   :      0   Min.   :     0
##  1st Qu.:    79   1st Qu.:    93.0   1st Qu.:    419   1st Qu.:   229
##  Median :   142   Median :   155.0   Median :    865   Median :   474
##  Mean   :   595   Mean   :   588.4   Mean   :   2802   Mean   :  1150
##  3rd Qu.:   304   3rd Qu.:   305.0   3rd Qu.:   1816   3rd Qu.:  1108
##  Max.   :675738   Max.   :521716.0   Max.   :1252320   Max.   :175719
##      Width             Time
##  Min.   :  22.0   Min.   :  137.0
##  1st Qu.:  46.0   1st Qu.:  274.0
##  Median :  55.0   Median :  405.0
##  Mean   :  60.3   Mean   :  425.6
##  3rd Qu.:  63.0   3rd Qu.:  530.0
##  Max.   :1001.0   Max.   :43123.0
```

Dataset 4:

```
file_list4 <- list.files(path="../FACS_Data/PHBV_Micha4/")
dataset4 <- data.frame()
for(i in 1:length(file_list4)){
  file_string <- file_list4[i]
  file_string <- paste("../FACS_Data/PHBV_Micha4/", file_string, sep = "")
  fcs.data <- read.FCS(file_string)
  fcs.data.frame <- as.data.frame(exprs(fcs.data))
  dataset4 <- rbind(dataset4, fcs.data.frame)
}
head(dataset4)
```

```
##     FSC-A  SSC-A FL1-A FL2-A FL3-A FL4-A   FSC-H  SSC-H FL1-H FL2-H FL3-H FL4-H
## 1  166439  17010    18     0    55   138  122176  17492    25    33   114   105
## 2  162626 198662    52    20    13   243  122026 164360    37    50    47    24
## 3  132246  17775     0    33     0   238   83627  15705    11    51    16    78
## 4 2643012 399071  2811  2134  8352  1929 2155125 441919  2159  1502  7846  1847
## 5 2645515 632566    76   124   109   596 2153855 737383    60    49   116   372
## 6  152061 311247    30    15   293   206   82718 235704    45    88   202    28
##   Width Time
## 1    54  179
## 2    53  184
## 3    55  185
## 4    80  186
## 5    86  190
## 6    61  190
```

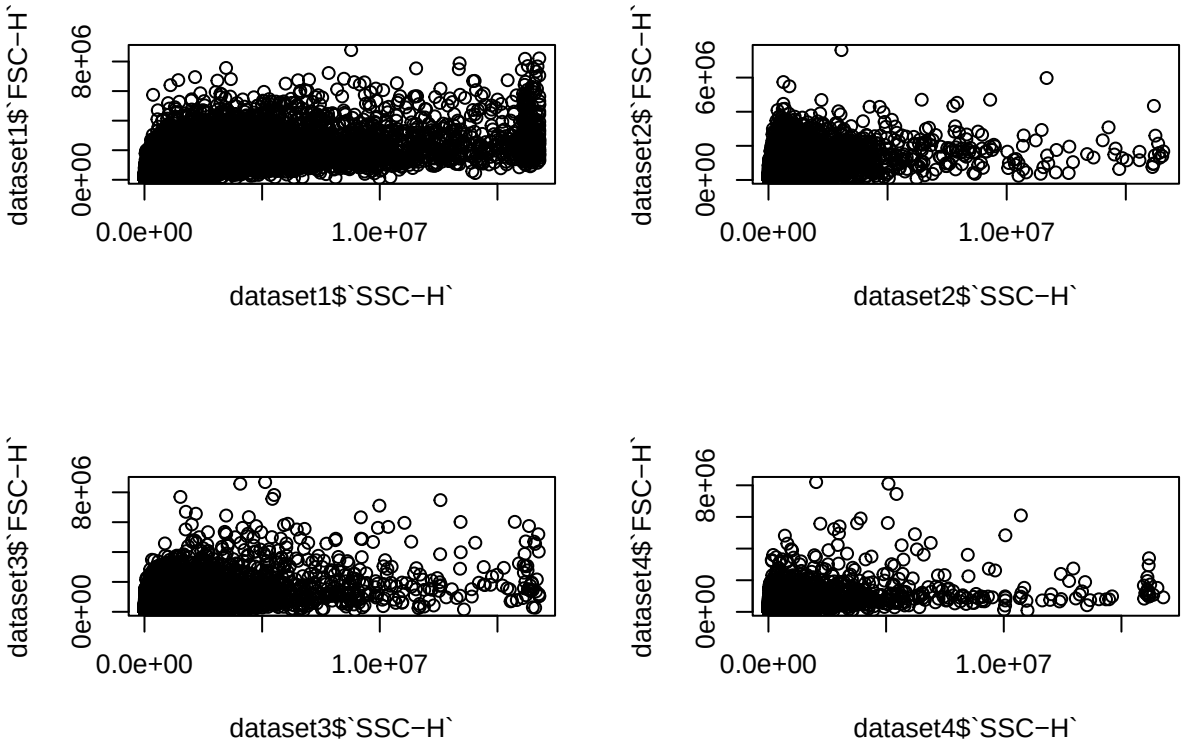
```
summary(dataset4)
```

```
##      FSC-A              SSC-A              FL1-A             FL2-A
##  Min.   :   54977   Min.   :       0   Min.   :      0   Min.   :      0
##  1st Qu.:  148896   1st Qu.:   26978   1st Qu.:     46   1st Qu.:     60
##  Median :  247566   Median :   68998   Median :    195   Median :    166
##  Mean   :  444931   Mean   :  348154   Mean   :   5047   Mean   :   2984
##  3rd Qu.:  484845   3rd Qu.:  204381   3rd Qu.:   1130   3rd Qu.:    704
##  Max.   :16777215   Max.   :16777215   Max.   :6626051   Max.   :4197795
##      FL3-A             FL4-A              FSC-H             SSC-H
##  Min.   :      0   Min.   :     0.0   Min.   :  80000   Min.   :    1263
##  1st Qu.:    117   1st Qu.:   249.0   1st Qu.: 124400   1st Qu.:   32854
##  Median :    405   Median :   339.0   Median : 208978   Median :   80768
##  Mean   :   5806   Mean   :   858.5   Mean   : 338792   Mean   :  333388
##  3rd Qu.:   1654   3rd Qu.:   538.0   3rd Qu.: 414710   3rd Qu.:  213178
##  Max.   :6427827   Max.   :867614.0   Max.   :8201308   Max.   :16777215
##      FL1-H             FL2-H             FL3-H              FL4-H
##  Min.   :      0   Min.   :     0   Min.   :      0.0   Min.   :     0
##  1st Qu.:     55   1st Qu.:    71   1st Qu.:    144.8   1st Qu.:    97
##  Median :    168   Median :   140   Median :    393.0   Median :   237
##  Mean   :   2890   Mean   :  1632   Mean   :   3431.0   Mean   :   672
##  3rd Qu.:    965   3rd Qu.:   578   3rd Qu.:   1477.0   3rd Qu.:   559
##  Max.   :1378027   Max.   :863503   Max.   :1195314.0   Max.   :207005
##      Width             Time
##  Min.   : 20.00   Min.   : 143.0
##  1st Qu.: 44.00   1st Qu.: 269.0
##  Median : 52.00   Median : 385.0
##  Mean   : 57.21   Mean   : 514.4
##  3rd Qu.: 61.00   3rd Qu.: 499.0
##  Max.   :937.00   Max.   :5837.0
```

Visualtization

For the first steps of the analysis hierarchical clusters will be made to figure out how many clusters exist in the datasets, as each of these clusters could represent a plastic or a bacteria specie.

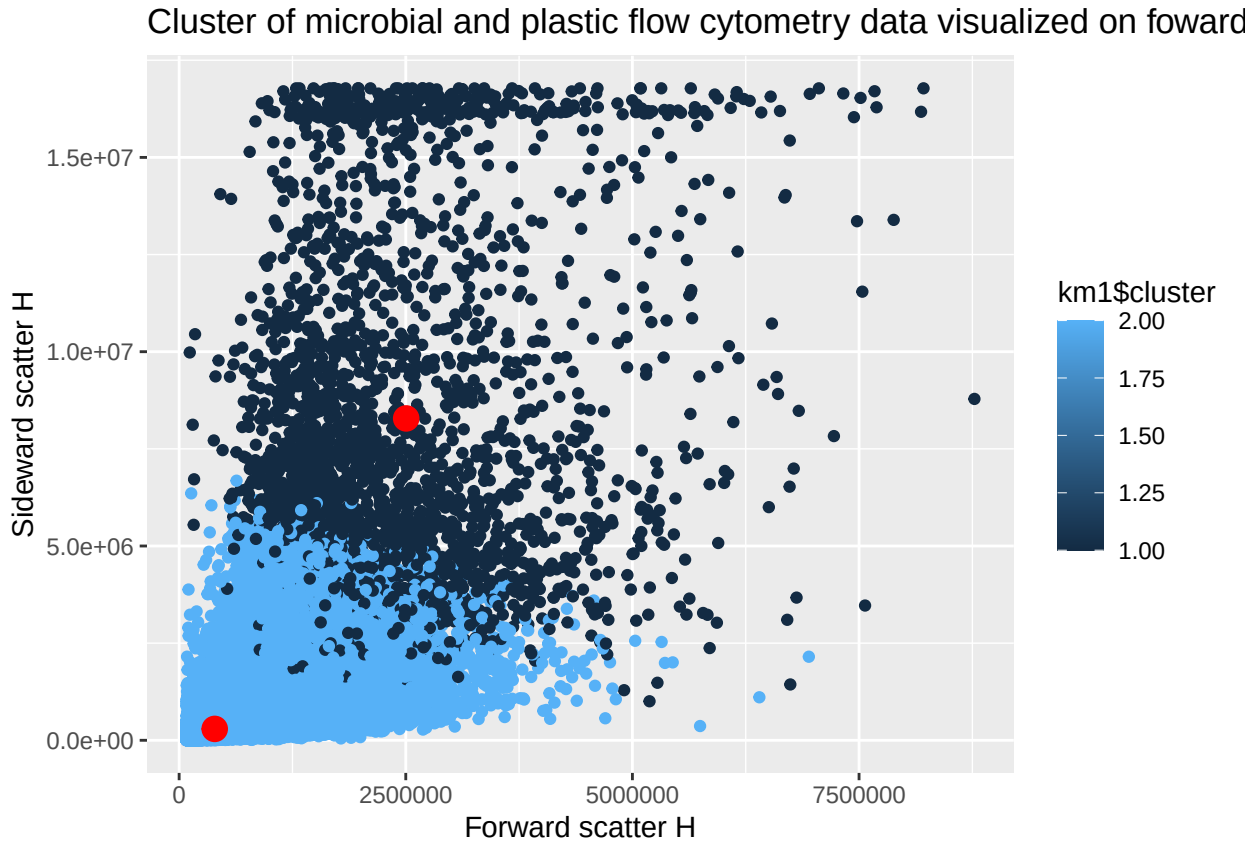
```
par(mfrow=c(2,2))
plot(dataset1$`FSC-H` ~ dataset1$`SSC-H`)
plot(dataset2$`FSC-H` ~ dataset2$`SSC-H`)
plot(dataset3$`FSC-H` ~ dataset3$`SSC-H`)
plot(dataset4$`FSC-H` ~ dataset4$`SSC-H`)
```



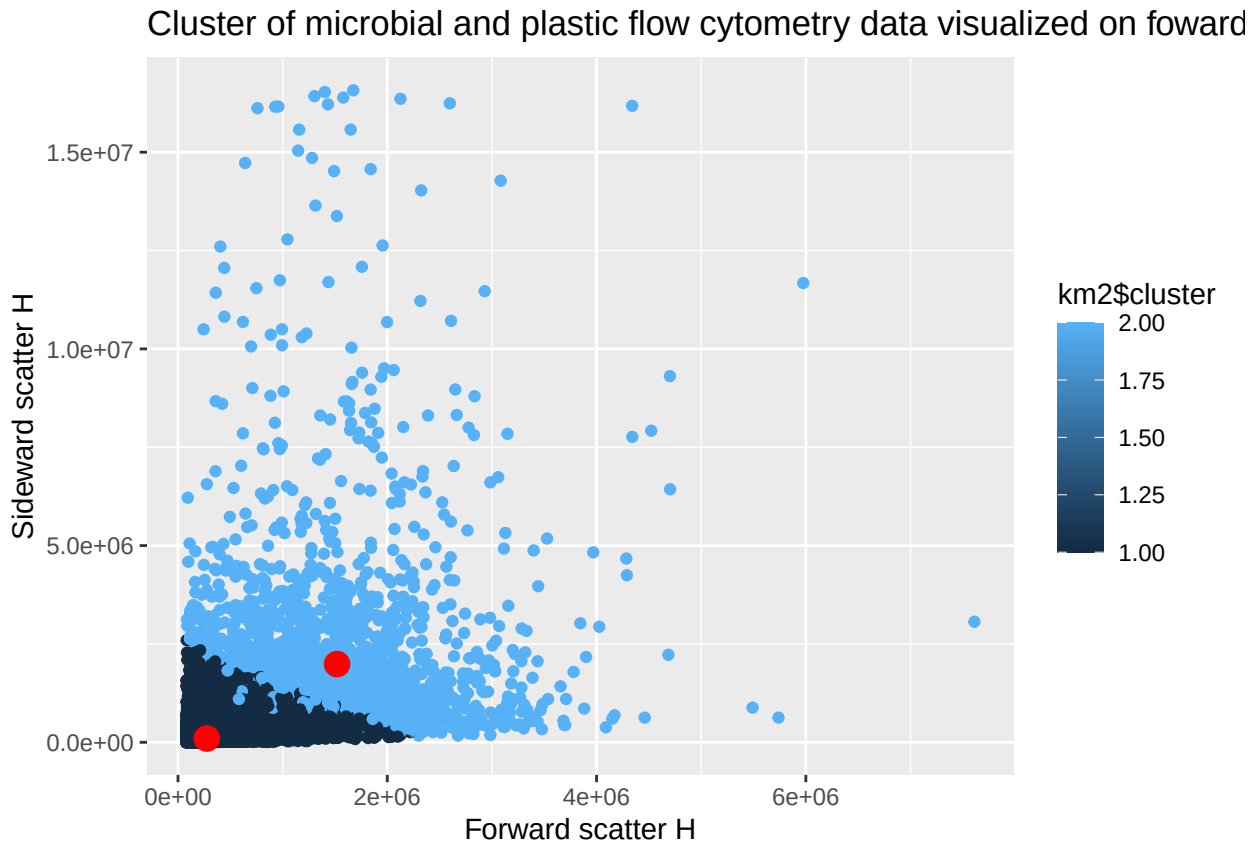
Clustering

Cluster data plotted using height scatter.

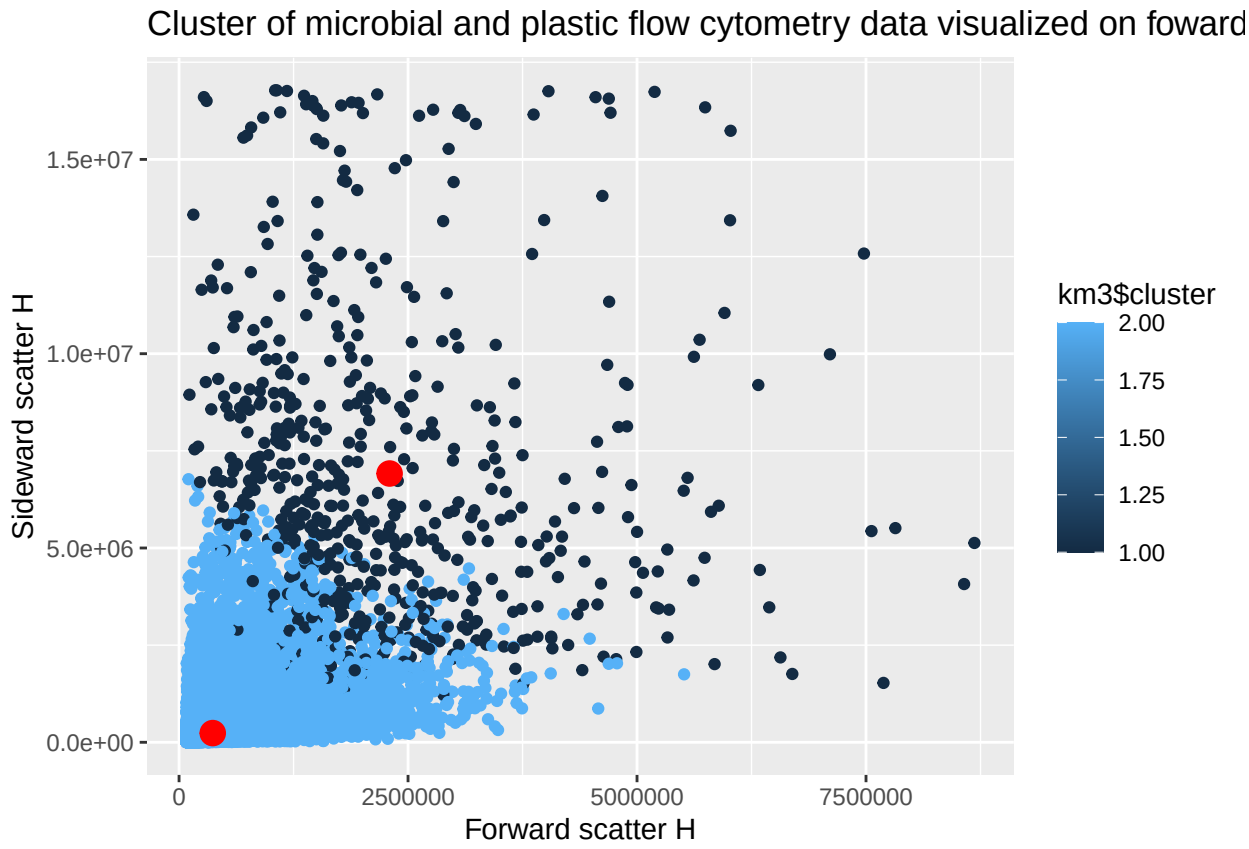
```
par(mfrow=c(2,2))
km1 <- kmeans(dataset1[,c(1:12)], 2)
km2 <- kmeans(dataset2[,c(1:12)], 2)
km3 <- kmeans(dataset3[,c(1:12)], 2)
km4 <- kmeans(dataset4[,c(1:12)], 2)
ggplot() + geom_point(data = dataset1, mapping = aes(x = `FSC-H`, y = `SSC-H`, colour = km1$cluster)) +
  geom_point(mapping = aes(x = km1$centers[,7], y = km1$centers[,8]), color = "red", size = 4) +
  ggtitle("Cluster of microbial and plastic flow cytometry data visualized on foward vs sideward height
  xlab("Forward scatter H") + ylab("Sideward scatter H")
```



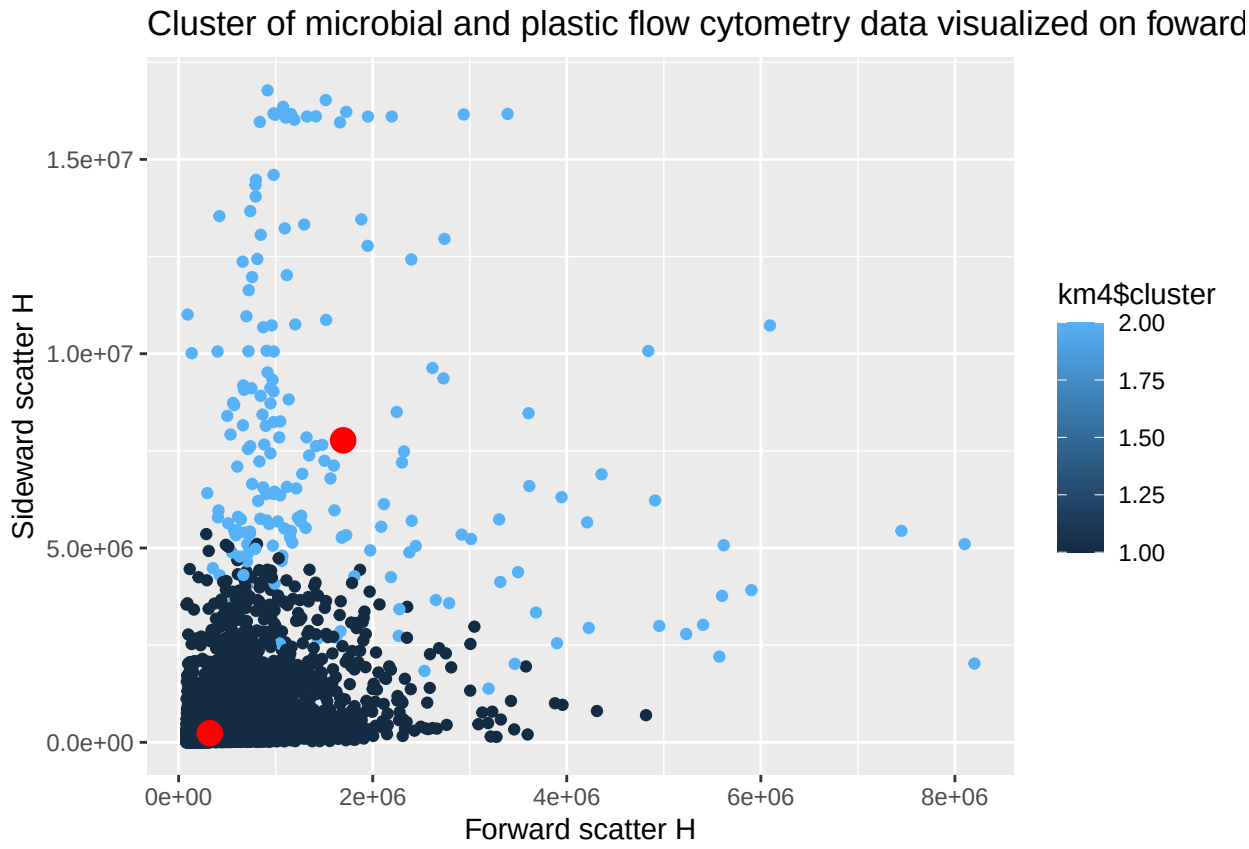

```
ggplot() + geom_point(data = dataset2, mapping = aes(x = `FSC-H`, y = `SSC-H`, colour = km2$cluster)) +
  geom_point(mapping = aes(x = km2$centers[,7], y = km2$centers[,8]), color = "red", size = 4) +
  ggtitle("Cluster of microbial and plastic flow cytometry data visualized on foward vs sideward height
  xlab("Forward scatter H") + ylab("Sideward scatter H")
```




```
ggplot() + geom_point(data = dataset3, mapping = aes(x = `FSC-H`, y = `SSC-H`, colour = km3$cluster)) +
  geom_point(mapping = aes(x = km3$centers[,7], y = km3$centers[,8]), color = "red", size = 4) +
  ggtitle("Cluster of microbial and plastic flow cytometry data visualized on foward vs sideward height
  xlab("Forward scatter H") + ylab("Sideward scatter H")
```

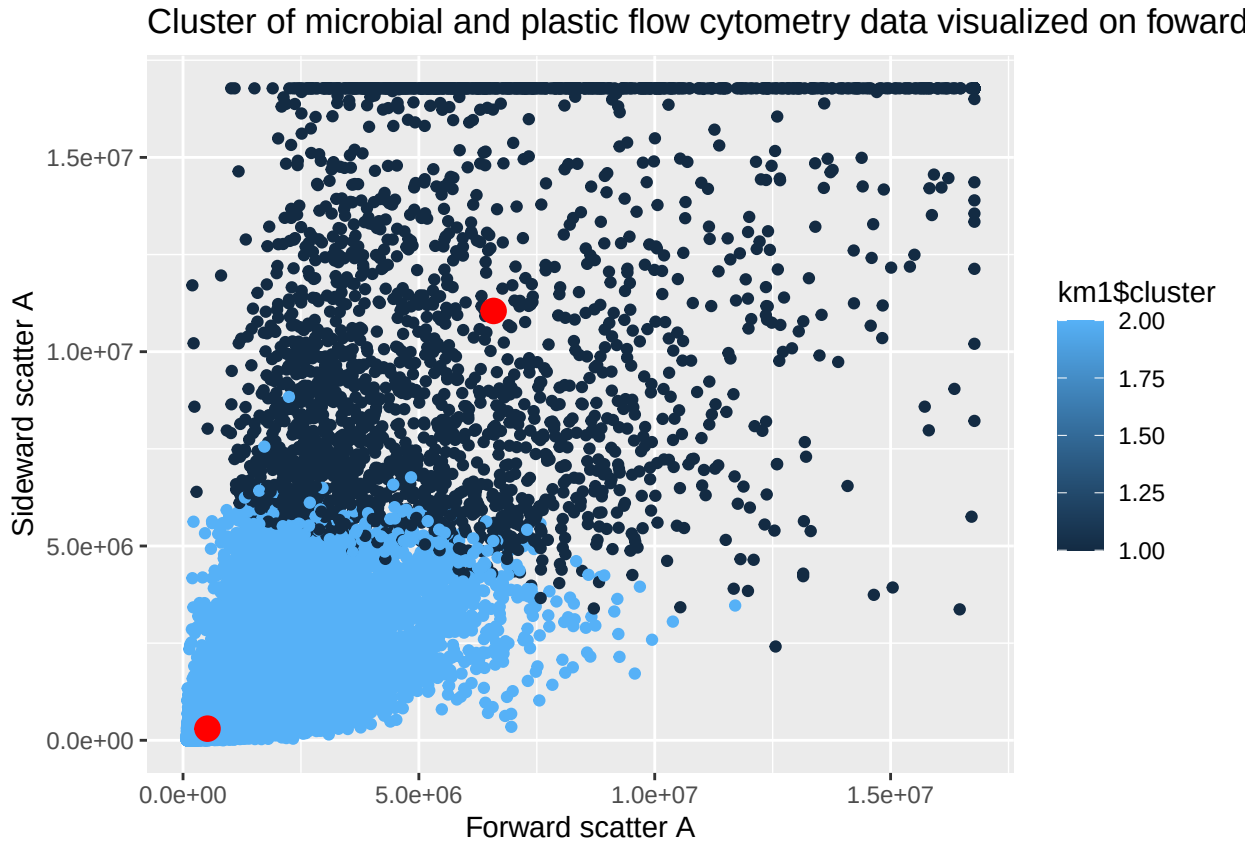



```
ggplot() + geom_point(data = dataset4, mapping = aes(x = `FSC-H`, y = `SSC-H`, colour = km4$cluster)) +
  geom_point(mapping = aes(x = km4$centers[,7], y = km4$centers[,8]), color = "red", size = 4) +
  ggtitle("Cluster of microbial and plastic flow cytometry data visualized on foward vs sideward height
  xlab("Forward scatter H") + ylab("Sideward scatter H")
```

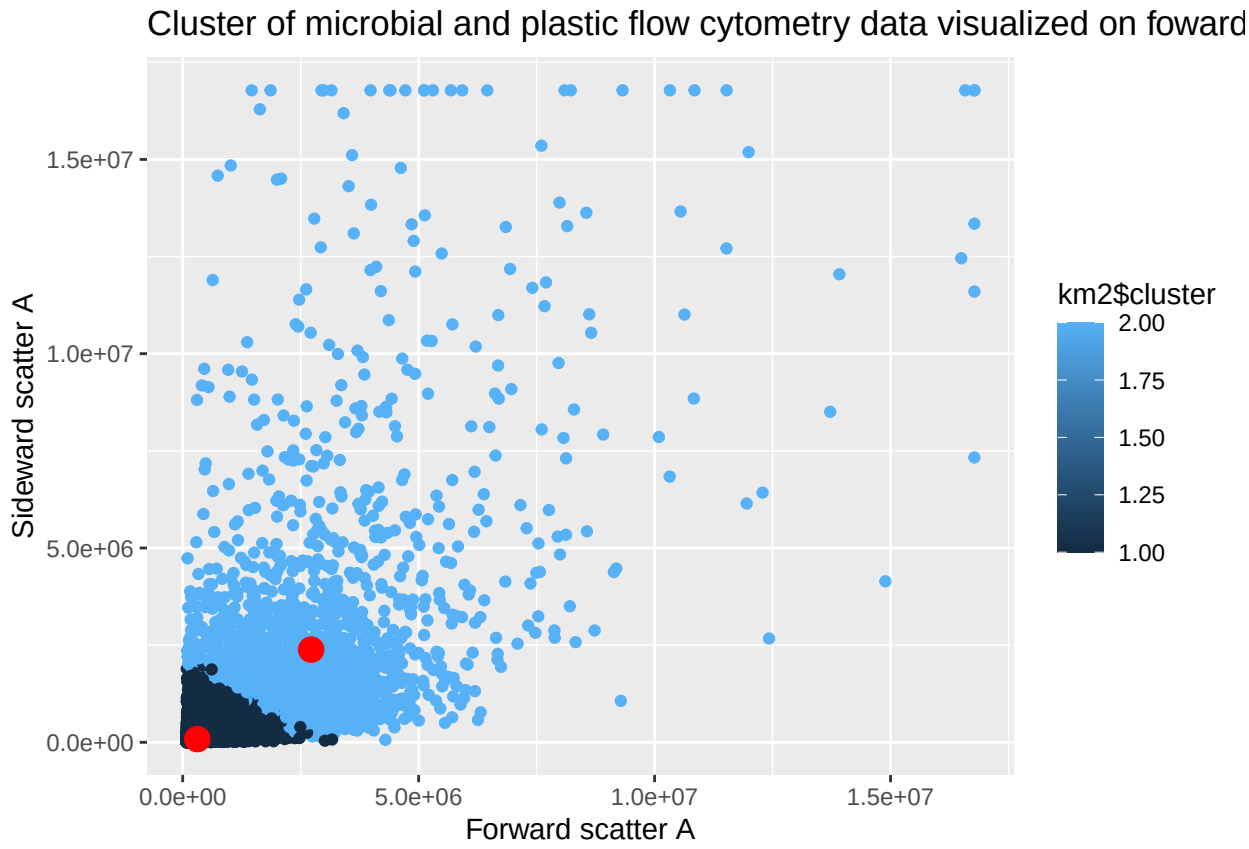


Cluster data using area scatter.

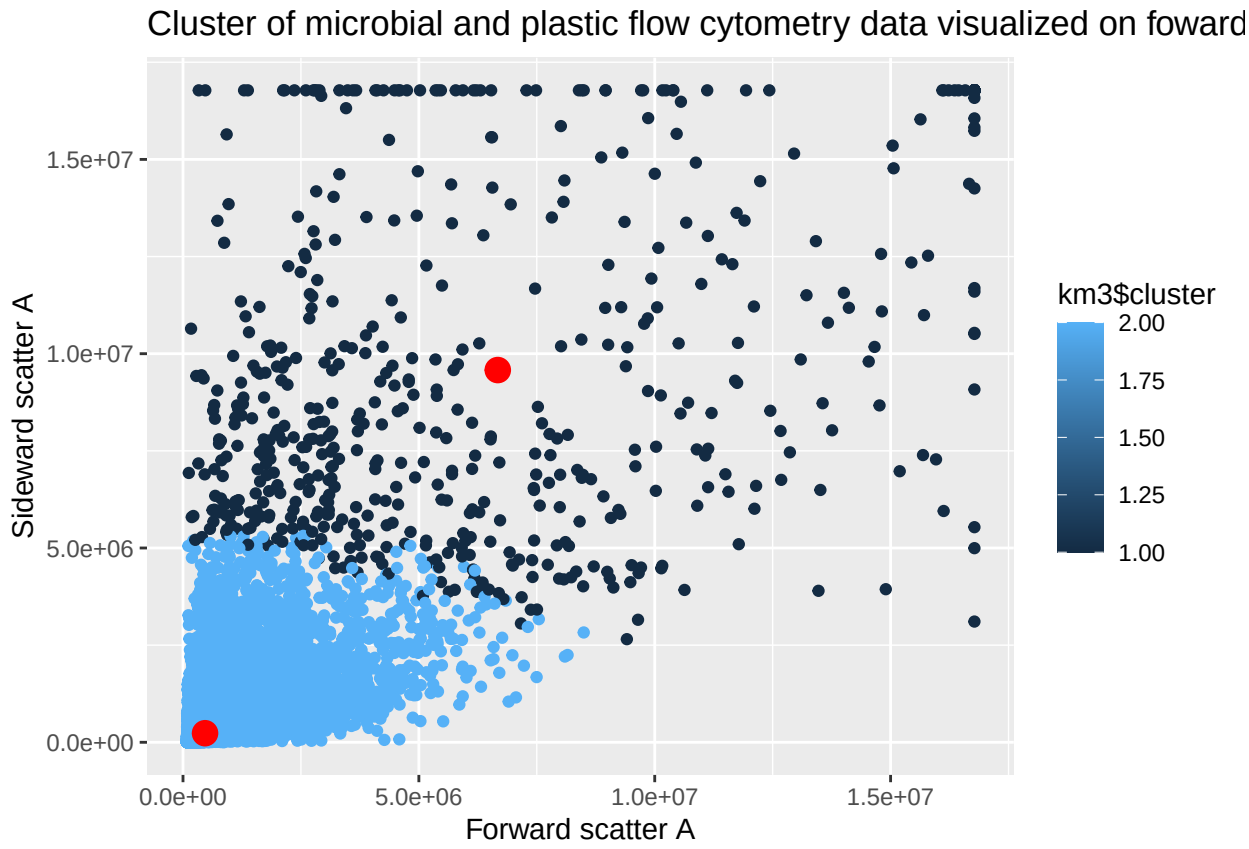
```
par(mfrow=c(2,2))
ggplot() + geom_point(data = dataset1, mapping = aes(x = `FSC-A`, y = `SSC-A`, colour = km1$cluster)) +
  geom_point(mapping = aes(x = km1$centers[,1], y = km1$centers[,2]), color = "red", size = 4) +
  ggtitle("Cluster of microbial and plastic flow cytometry data visualized on foward vs sideward area s
  xlab("Forward scatter A") + ylab("Sideward scatter A")
```



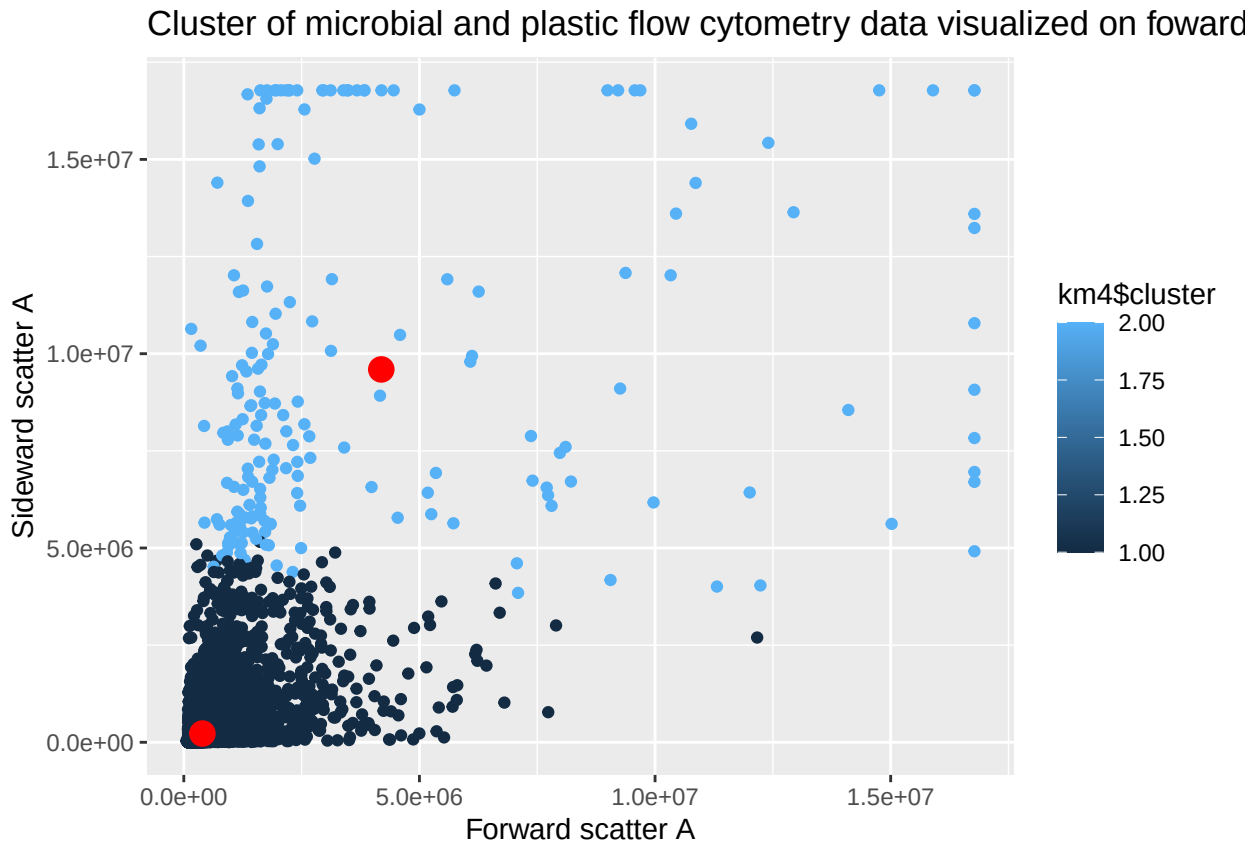

```
ggplot() + geom_point(data = dataset2, mapping = aes(x = `FSC-A`, y = `SSC-A`, colour = km2$cluster)) +
  geom_point(mapping = aes(x = km2$centers[,1], y = km2$centers[,2]), color = "red", size = 4) +
  ggtitle("Cluster of microbial and plastic flow cytometry data visualized on foward vs sideward area s
  xlab("Forward scatter A") + ylab("Sideward scatter A")
```




```
ggplot() + geom_point(data = dataset3, mapping = aes(x = `FSC-A`, y = `SSC-A`, colour = km3$cluster)) +
  geom_point(mapping = aes(x = km3$centers[,1], y = km3$centers[,2]), color = "red", size = 4) +
  ggtitle("Cluster of microbial and plastic flow cytometry data visualized on foward vs sideward area s
  xlab("Forward scatter A") + ylab("Sideward scatter A")
```




```
ggplot() + geom_point(data = dataset4, mapping = aes(x = `FSC-A`, y = `SSC-A`, colour = km4$cluster)) +
  geom_point(mapping = aes(x = km4$centers[,1], y = km4$centers[,2]), color = "red", size = 4) +
  ggtitle("Cluster of microbial and plastic flow cytometry data visualized on foward vs sideward area s
  xlab("Forward scatter A") + ylab("Sideward scatter A")
```



As seen in the figures above the clusters are mostly separated by an increase in scattering, with the first cluster sitting in the bottom left of the plot while the second cluster starts forming after a sideward scatter of just before 5.0e+06 with the exception of the second dataset. A higher forward scatter also seems to resemble the second cluster but seems to be less important than the difference in sideward scatter.

To determine which cluster specifies which group the types of scatters have to be known. According to the fcs documentation forward scatter describes the size of the particle and sideward the complexity of the particle. Knowing this it can be assumed that the black cluster are the plastics and the red cluster the microbes as the microbes have a more complex surface with structures on their cell walls while the plastics should be relatively simple and the same form, meaning that they are centered more on the left of the plot.

Creating of csv files:

```
write.csv(dataset1, file = "./CSVdata/dataset1.csv", row.names = FALSE)
write.csv(dataset2, file = "./CSVdata/dataset2.csv", row.names = FALSE)
write.csv(dataset3, file = "./CSVdata/dataset3.csv", row.names = FALSE)
write.csv(dataset4, file = "./CSVdata/dataset4.csv", row.names = FALSE)
```

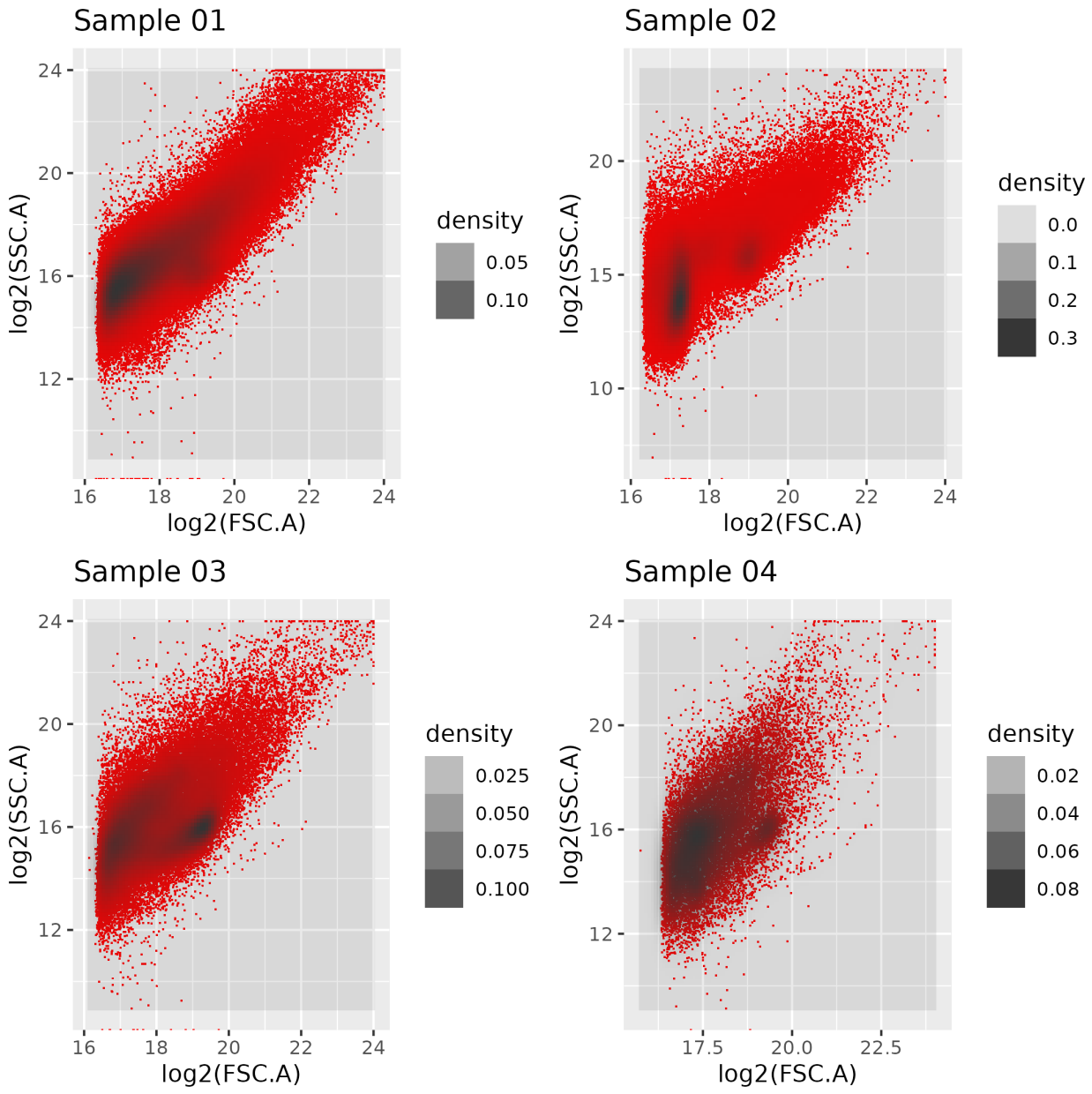


Figure 1: Density plot

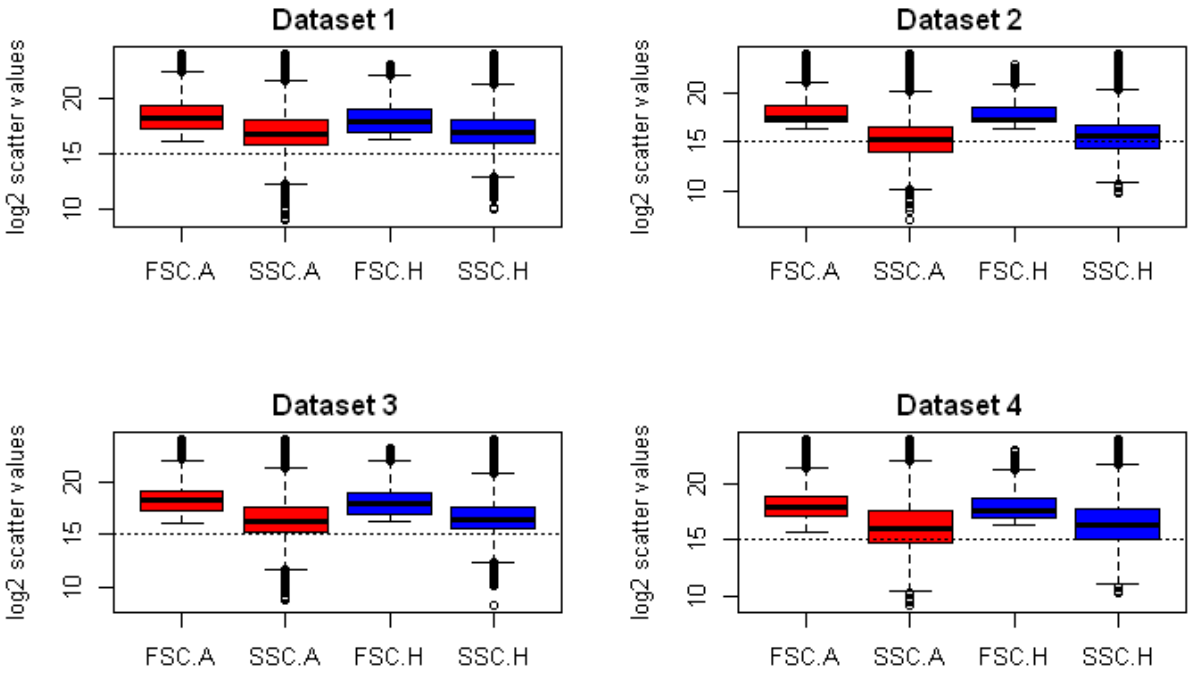


Figure 2: Boxplot of the scatter values.