

# Profil\_4cer\_preprocessing

Asger\_Wretlind

4/2/2022

```
#import raw ms peaks from MZmine step 12 export  
data_raw_peaks <- vroom::vroom(here::here("data-raw/0033_profil_4cer_export_040222.csv"))
```

```
## New names:  
## Rows: 1942 Columns: 948  
## -- Column specification  
## ----- Delimiter: ";" dbl  
## (946): row ID, row m/z, row retention time, 0033_LIP1p_20170412_006_CALI... lgl  
## (2): row identity, ...948  
## i Use `spec()` to retrieve the full column specification for this data. i  
## Specify the column types or set `show_col_types = FALSE` to quiet this message.  
## * `` -> `...948`
```

```
# #View data to check if it is correctly loaded  
# View(data_raw_peaks)
```

```
#Select peaks based on specific IDs  
#input peak ID based on viewed chromatograms
```

```
ID_list <- c("Target_17_H_ISTD" = 9,  
            "Target_17_H20_ISTD" = 5,  
            "Target_16_H" = 1207,  
            "Target_16_H20" = 471,  
            "Target_18_H" = 1079,  
            "Target_18_H20" = 1022,  
            "Target_20_H" = 1049,  
            "Target_20_H20" = 1000,  
            "Target_22_H" = 964,  
            "Target_22_H20" = 714,  
            "Target_24_0_H" = 328,  
            "Target_24_0_H20" = 229,  
            "Target_24_1_H" = 959,  
            "Target_24_1_H20" = 954)
```

```
ID_list <- ID_list[order(ID_list)]
```

```
data <- data_raw_peaks %>%  
  filter(data_raw_peaks$`row ID` %in% ID_list) %>%  
  arrange(`row ID`) %>%  
  mutate(`row identity` = names(ID_list)) %>%  
  arrange(`row identity`)
```

```
rm(ID_list, data_raw_peaks)
```

```
#NOTE This is project specific cleaning
```

```
#Remove weird artifact at the final data column
```

```
data <- data[,-length(data)]
```

```
#Keep only H2o adducts, remove RT, mz, row ID a and pivot table
```

```
data <- data %>%  
  filter(grepl("H2O", `row identity`)) %>%  
  mutate(ID = gsub("_H2O", "", `row identity`)) %>%  
  select(-c("row m/z", "row retention time", "row identity", "row ID")) %>%  
  pivot_longer(cols = -ID, names_to = "Steno ID") %>%  
  pivot_wider(names_from = ID, values_from = value)
```

```
#Separate Steno ID into a new columns
```

```
data <- data %>%  
  separate(`Steno ID`,  
    c("Project_nr",  
      "Method",  
      "Date",  
      "Run_nr",  
      "Sample_ID",  
      "Repeat_A",  
      "Repeat_B"))
```

```
## Warning: Expected 7 pieces. Additional pieces discarded in 943 rows [1, 2, 3, 4,  
## 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
#Run plot for
```

```
Run_nr_plot <- function(Data, Target, Filter_out = FALSE) {  
  Target_pos <- which(colnames(Data) %in% Target)  
  Target <- sym(Target)  
  
  tmp_bounds <- data.frame("Mean" = NA, "SD" = NA, "lower" = NA, "upper" = NA)  
  
  Data %>%  
    filter(!grepl(Filter_out, `Sample_ID`)) %>%  
    pull(Target) %>%  
    mean() -> tmp_bounds$Mean  
  
  Data %>%  
    filter(!grepl(Filter_out, `Sample_ID`)) %>%  
    pull(Target) %>%  
    sd() -> tmp_bounds$SD  
  
  tmp_bounds$lower <- tmp_bounds$Mean - 2*tmp_bounds$SD  
  tmp_bounds$upper <- tmp_bounds$Mean + 2*tmp_bounds$SD  
  
  p <- Data %>%  
    filter(!grepl(Filter_out, `Sample_ID`)) %>%
```

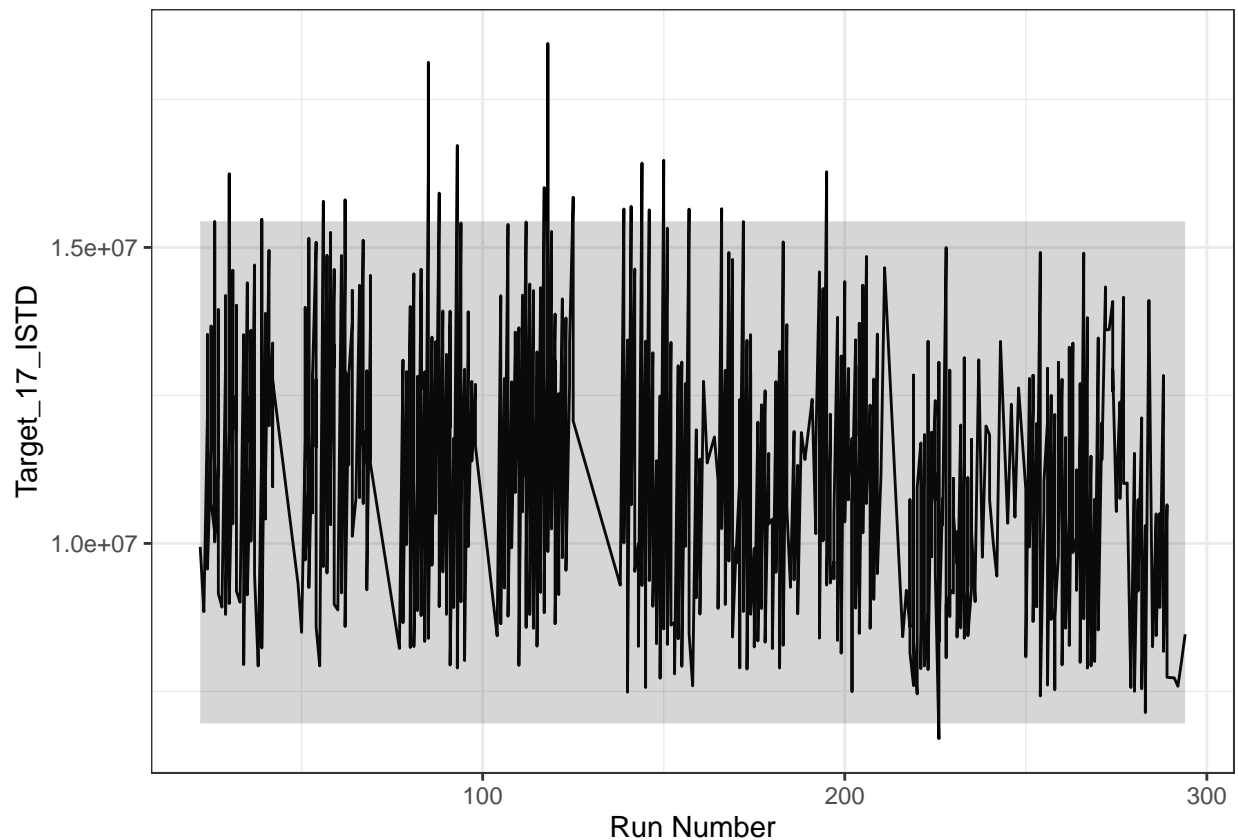
```

  ggplot(aes(x = as.numeric(Run_nr), y = !!Target)) +
    geom_line() +
    geom_ribbon(aes(ymin = tmp_bounds$lower, ymax = tmp_bounds$upper),
              alpha = 0.2) +
    xlab(label = "Run Number")+
    theme_bw()

  return(p)
}

#ISTD Run nr plot
Run_nr_plot(Data = data, Target = "Target_17_ISTD", Filter_out = "[A-Za-z]+")

```



```

#Relative Standard deviation (RSD) of Pooled samples before proxy conc. normalization
data %>%
  filter(Sample_ID == "P0") %>%
  summarise(across(starts_with("Target_"), ~ sd(.) / mean(.) * 100, .names = "RSD_{.col}"))

```

```

## # A tibble: 1 x 7
##   RSD_Target_16 RSD_Target_17_ISTD RSD_Target_18 RSD_T~1 RSD_T~2 RSD_T~3 RSD_T~4
##           <dbl>           <dbl>         <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1           17.5             17.5          27.9    30.0    32.7    31.6    31.1
## # ... with abbreviated variable names 1: RSD_Target_20, 2: RSD_Target_22,
## # 3: RSD_Target_24_0, 4: RSD_Target_24_1

```

```

#Proxy conc.
#ISTD concentration in each sample ends up at 2ug/ml
data <- data %>%
  mutate(across(.cols = starts_with("Target") & !contains("ISTD"),
    ~ ./(Target_17_ISTD*2)))

#Relative Standard deviation (RSD) of Pooled samples after proxy conc. normalization
data %>%
  filter(Sample_ID == "PO") %>%
  summarise(across(starts_with("Target_"), ~ sd(.) / mean(.) * 100, .names = "RSD_{.col}"))

```

```

## # A tibble: 1 x 7
##   RSD_Target_16 RSD_Target_17_ISTD RSD_Target_18 RSD_T~1 RSD_T~2 RSD_T~3 RSD_T~4
##         <dbl>         <dbl>         <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1         11.2         17.5         15.1    18.5    21.8    21.3    20.9
## # ... with abbreviated variable names 1: RSD_Target_20, 2: RSD_Target_22,
## #   3: RSD_Target_24_0, 4: RSD_Target_24_1

```

```

##NOTE Normalizing to ISTD should no be carried out if using proxy conc.

```

```

# #Normalization
# # #cer 24_1 Run nr plot before normalizing to ISTD
# # Run_nr_plot(Data = data, Target = "Target_24_1", Filter_out = "[A-Za-z]+")
#
# #Relative Standard deviation (RSD) of Pooled samples before normalization
# data %>%
#   filter(Sample_ID == "PO") %>%
#   summarise(across(starts_with("Target_"), ~ sd(.) / mean(.) * 100, .names = "RSD_{.col}"))
#
# #Normalize to ISTD
# data <- data %>%
#   mutate(across(starts_with("Target") & !contains("ISTD"),
#     ~ . / Target_17_ISTD))
#
# # #cer 24_1 Run nr plot before normalizing to ISTD
# # Run_nr_plot(Data = data, Target = "Target_24_1", Filter_out = "[A-Za-z]+")
#
# #Relative Standard deviation (RSD) of Pooled samples after normalization
# data %>%
#   filter(Sample_ID == "PO") %>%
#   summarise(across(starts_with("Target_"), ~ sd(.) / mean(.) * 100, .names = "RSD_{.col}"))
#

```

```

#Remove QC samples and unnecessary features

```

```

data <- data %>%
  filter(!grepl("[A-Za-z]+", `Sample_ID`)) %>%
  select(-c(Project_nr, Method, Date, Run_nr, Repeat_A, Repeat_B)) %>%
  select(-Target_17_ISTD)

```

```

#Ratios of interest (16:0)/(24:0), (18:0)/(24:0), (24:1)/(24:0)

```

```

#Calculate ratio of all targets against cer(24:0)
data_ratios <- data %>%
  mutate(across(contains("Target") & !contains("_24_0"),
    ~ . / Target_24_0)) %>%
  select(-Target_24_0) %>%
  rename(`Ratio 16/24:0` = Target_16) %>%
  rename(`Ratio 18/24:0` = Target_18) %>%
  rename(`Ratio 20/24:0` = Target_20) %>%
  rename(`Ratio 22/24:0` = Target_22) %>%
  rename(`Ratio 24:1/24:0` = Target_24_1)

data <- data %>%
  left_join(x = ., y = data_ratios, by = "Sample_ID")

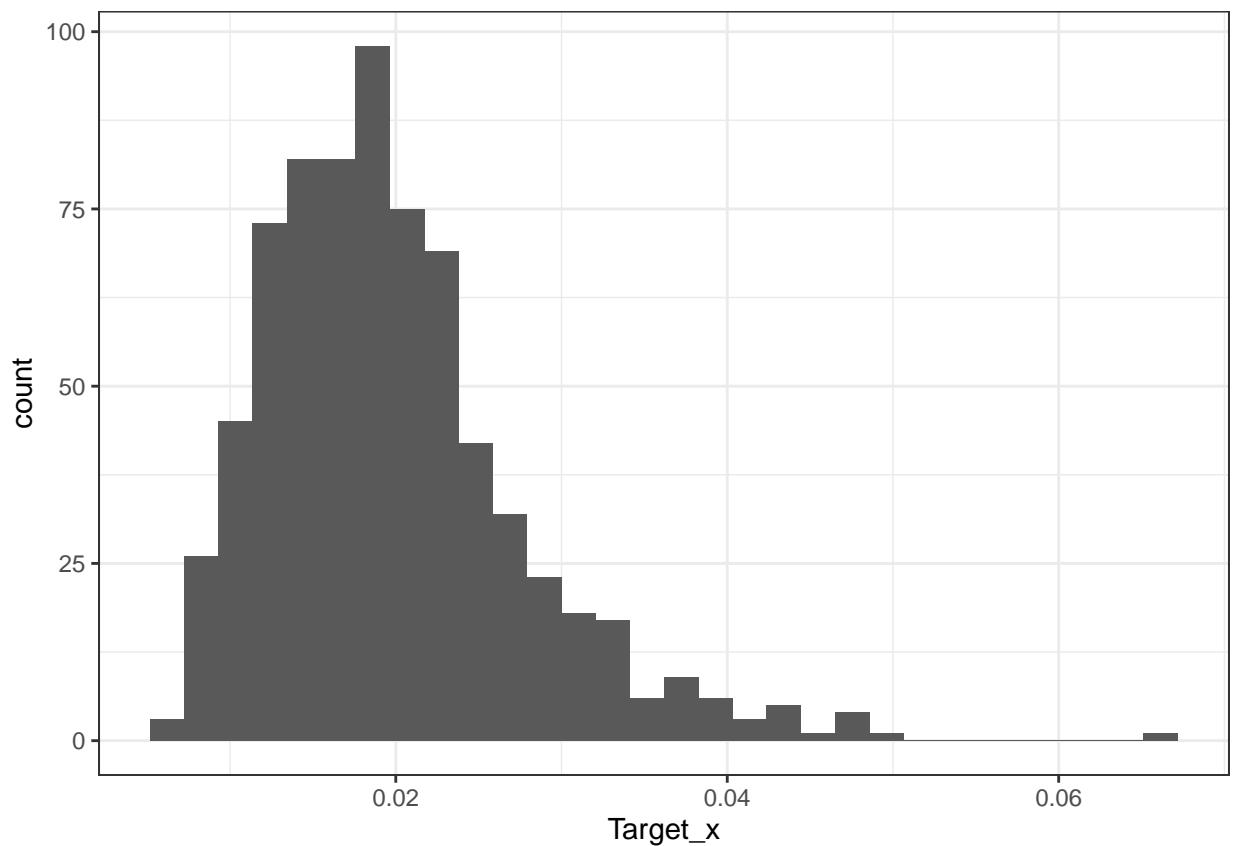
rm(data_ratios)

```

```

#plot distribution
data %>%
  mutate(Target_x = Target_24_1) %>%
  ggplot(aes(x = Target_x)) +
  geom_histogram(bins = 30) +
  theme_bw()

```

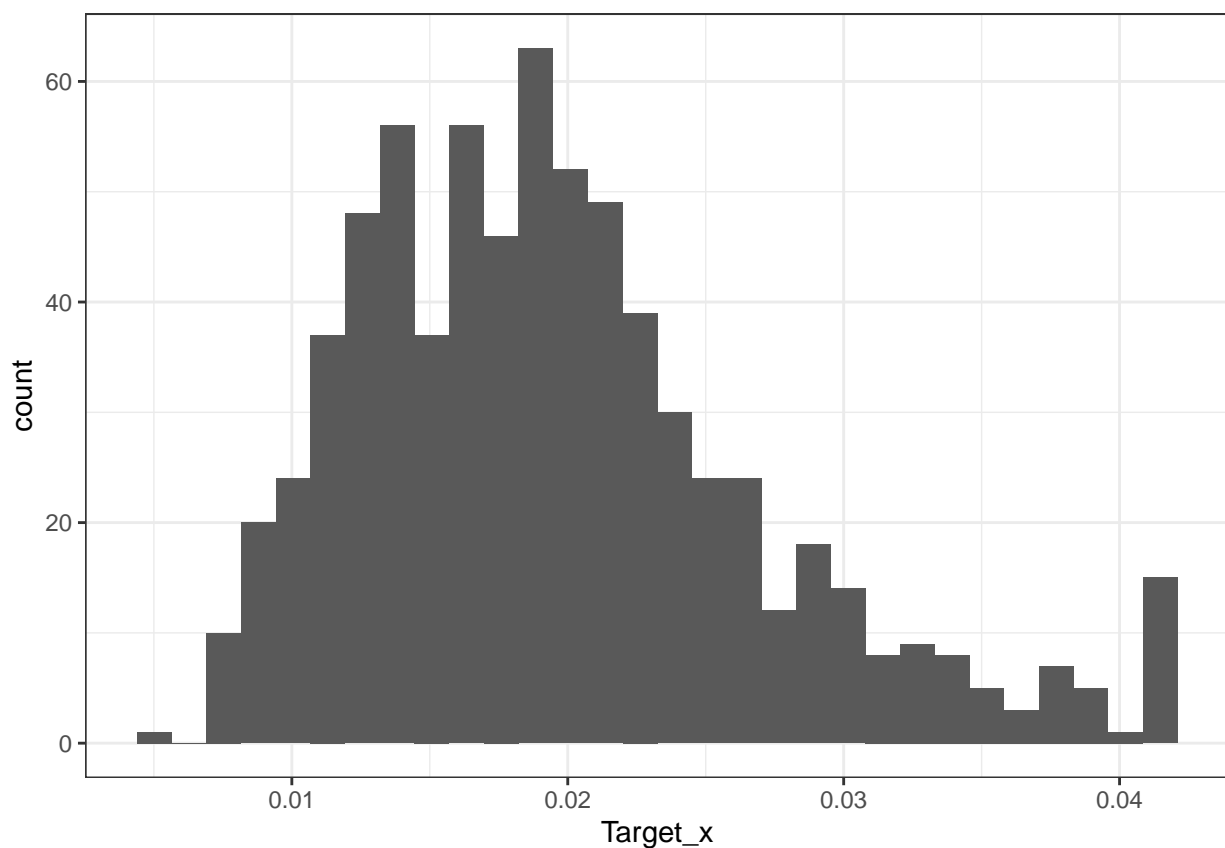


```

#Truncate outliers (outside median +- 3*sd) to median +- 3*sd
data <- data %>%
  mutate(across(starts_with("Target") | starts_with("Ratio"),
    ~ ifelse(. > median(.)+3*sd(.), median(.)+3*sd(.), .))) %>%
  mutate(across(starts_with("Target") | starts_with("Ratio"),
    ~ ifelse(. < median(.)-3*sd(.), median(.)-3*sd(.), .)))

#plot distribution
data %>%
  mutate(Target_x = Target_24_1 ) %>%
  ggplot(aes(x = Target_x)) +
  geom_histogram(bins = 30) +
  theme_bw()

```



```

#Import a curated set of the clinical variables - 667 observations
data_raw_clinical_curated <-
  vroom::vroom(file = here::here("data-raw/profil_selected_clinical_220322.csv")) %>%
  mutate(Curated = 1)

## New names:
## Rows: 667 Columns: 63
## -- Column specification
## ----- Delimiter: "," dbl
## (61): ...1, cpr, id_profil, Age, Duration_DM, Gender, Smoking, Weight, ... date
## (2): pro_date_index, pro_date_end

```



```

relocate(contains("cv_komb_profil"), .after = "Spiron") %>%
relocate(contains("doed"), .after = "Spiron") %>%
relocate(Blood_TGA, .after = Blood_CREAE)

#Correct Censor data
#The Surv function expect: 0=right censored (no event), 1=event at time, 2=left censored (death)
#some variables have this setup 0=event, 1=death, 2=no event/migration and needs correction
#censor_cv_komb_profil, censor_alb_prog, censor_gfrfald30_p, censor_ESRD_profil
data_clinical <- data_clinical %>%
  mutate(across(
    .cols = c(censor_cv_komb_profil, censor_alb_prog, censor_gfrfald30_p, censor_ESRD_profil),
    ~ . + 1)) %>%
  mutate(across(
    .cols = c(censor_cv_komb_profil, censor_alb_prog, censor_gfrfald30_p, censor_ESRD_profil),
    ~ ifelse(. == 3, 0, .)))

#The Surv function expect:0=no event, 1=event
#Some variables are recorded opposite and needs to be corrected that is:
#censor_doed_profil, censor_komb_nyre_endepunkt_p
data_clinical <- data_clinical %>%
  mutate(across(
    .cols = c(censor_doed_profil, censor_komb_nyre_endepunkt_p),
    ~ ifelse(. == 0, 1, 0)))

rm(data_raw_clinical_curated, data_raw_clinical_all, data_raw_clinical_missing)

#Merge clinical data and peak data
data <- data %>%
  mutate(Sample_ID = as.numeric(Sample_ID)) %>%
  left_join(., data_clinical, by = "Sample_ID")

rm(data_clinical)

#Export data
vroom::vroom_write(data, here::here("data/0033_profil_4cer_data_preprocessed.csv"))

```