Formative Assessment 2
Exploratory Data Analysis

Zion John Yousef T. Ramilo
February 14, 2025

For the first set of questions, we will look again at the CyTOF data Download CyTOF data. Each row in the dataset represents a cell, and each column in the dataset represents a protein, and the value is element i, j of the dataset represents the amount of protein j in cell i.

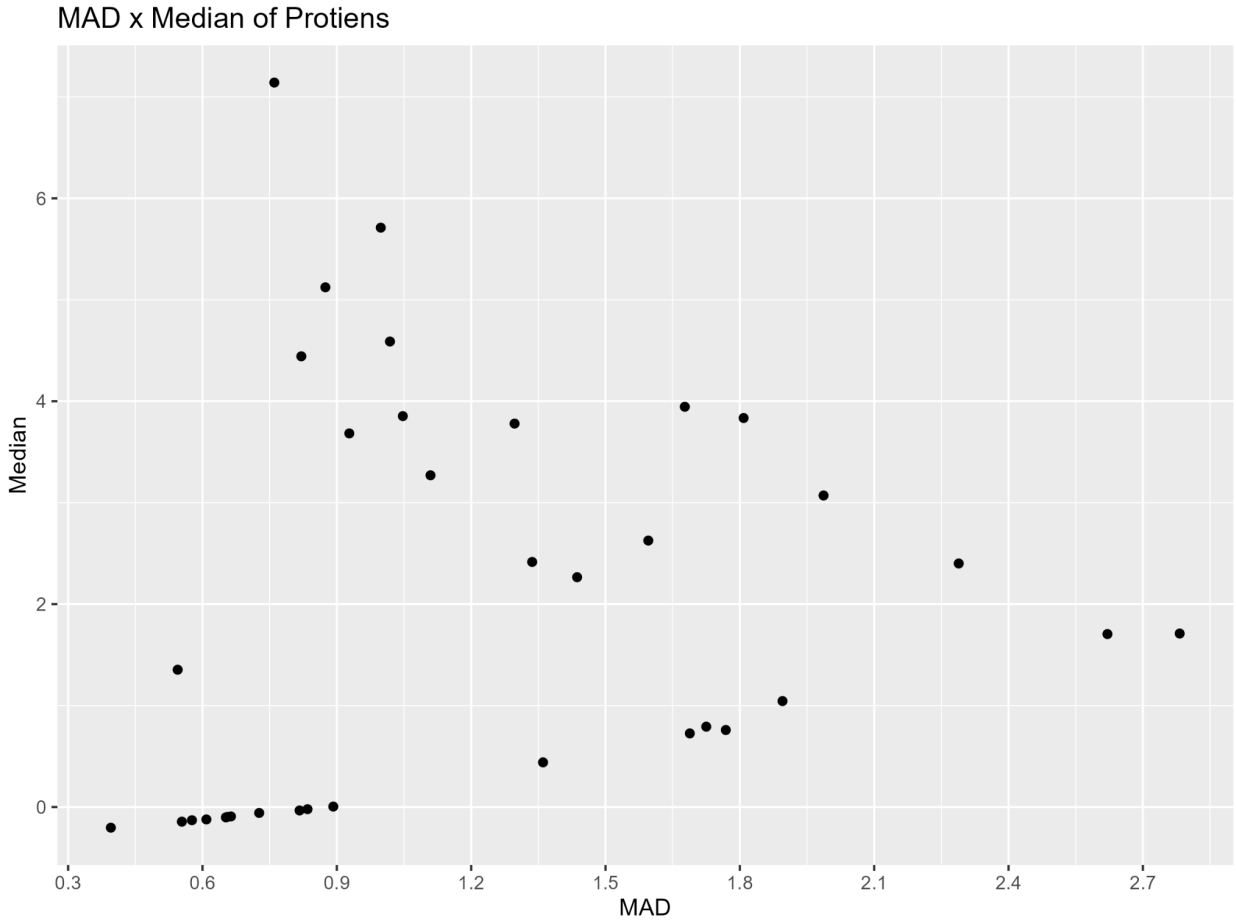
1. Use pivot_longer to reshape the dataset into one that has two columns, the first giving the protein identity and the second giving the amount of the protein in one of the cells. The dataset you get should have 1750000 rows (50000 cells in the original dataset times 35 proteins).

```
cytof_one_experiment_the_long_way <-
pivot_longer(cytof_one_experiment,cols = NKp30:INFg,names_to =
"Protien", values_to = "Amount")

str(cytof_one_experiment_the_long_way)

>>> tibble [1,750,000 × 2] (S3: tbl_df/tbl/data.frame)
 $ Protien: chr [1:1750000] "NKp30" "KIR3DL1" "NKp44" "KIR2DL1" ...
 $ Amount : num [1:1750000] 0.188 3.616 -0.561 -0.294 2.478 ...
```

2. Use group_by and summarise to find the median protein level and the median absolute deviation of the protein level for each marker. (Use the R functions median and mad).

```
median_MAD_Data <- cytof_one_experiment_the_long_way %>%
  group_by(Protien) %>%
  summarize(
    "Median" = median(Amount),
    "Median Absolute Deviation" = mad(Amount,center = median(Amount))
  )

>>> median_MAD_Data
# A tibble: 35 × 3
   Protien  Median `Median Absolute Deviation`
   <chr>     <dbl>                       <dbl>
 1 CD107a  -0.122                       0.609
 2 CD16     5.12                        0.874
 3 CD161    0.726                       1.69
 4 CD2      3.95                        1.68
 5 CD4     -0.204                       0.395
 6 CD56     5.71                        0.998
 7 CD57     3.07                        1.99
 8 CD69     4.59                        1.02
 9 CD8      2.40                        2.29
10 CXCR6   -0.0581                      0.727
# ℹ 25 more rows
# ℹ Use `print(n = ...)` to see more rows
```

3. Make a plot with the MAD on the x-axis and the median on the y-axis. This is known as a spreadlocation (s-l) plot. What does it tell you about the relationship between the median and the mad?



The higher the median absolute deviation the more the median of the proteins cluster together. Since MAD is about the deviation of the data it follows that if the data produces a high MAD the more the data is clustered around the median.

4. Using either pivot_longer on its own or pivot_longer in combination with separate, reshape the dataset so that it has columns for country, event, year, and score.

```
library(dcldata)
data(example_gymnastics_2)
View(example_gymnastics_2)
gymnasticsDataset2 <-
pivot_longer(example_gymnastics_2,cols=vault_2012:floor_2016,names_to =
"Event", values_to = "Score") %>%
  separate(Event,into=c("Event","Year"), sep="_")
View(gymnasticsDataset2)

>>> gymnasticsDataset2
# A tibble: 12 × 4
   country       Event Year  Score
   <chr>         <chr> <chr> <dbl>
 1 United States vault 2012   48.1
 2 United States floor 2012   45.4
 3 United States vault 2016   46.9
 4 United States floor 2016   46.0
 5 Russia        vault 2012   46.4
 6 Russia        floor 2012   41.6
 7 Russia        vault 2016   45.7
 8 Russia        floor 2016   42.0
 9 China         vault 2012   44.3
10 China         floor 2012   40.8
11 China         vault 2016   44.3
12 China         floor 2016   42.1
```