Question 1

```
library(readr)
library(ggplot2)
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
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
raw_data <- read_csv("~/pstat 197a/module1-group-11/data/biomarker-raw.csv")
```

```
## Rows: 156 Columns: 1320
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1319): Group, Target Full Name, E3 ubiquitin-protein ligase CHIP, CCAAT...
## dbl    (1): Protein 4.1
##
## 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.
```

```
head(raw_data)
```

```
## # A tibble: 6 x 1,320
##   Group `Target Full Name` E3 ubiquitin-protein ligase ~1 CCAAT/enhancer-bindi~2
##   <chr> <chr>              <chr>                          <chr>
## 1 <NA>  Target             CHIP                           CEBPB
## 2 ASD   <NA>               618.6                          1489.3
## 3 ASD   <NA>               512.2                          1697.8
## 4 ASD   <NA>               438.5                          1121.7
```

```
## 5 ASD   <NA>              505                           1209.7
## 6 ASD   <NA>              440.7                         1120.2
## # i abbreviated names: 1: 'E3 ubiquitin-protein ligase CHIP',
## #   2: 'CCAAT/enhancer-binding protein beta'
## # i 1,316 more variables: 'Gamma-enolase' <chr>,
## #   'E3 SUMO-protein ligase PIAS4' <chr>,
## #   'Interleukin-10 receptor subunit alpha' <chr>,
## #   'Signal transducer and activator of transcription 3' <chr>,
## #   'Interferon regulatory factor 1' <chr>, ...
```

```
nrow(raw_data)
```

```
## [1] 156
```

```
raw_data <- raw_data[-1,] %>%
  select(-2, -8:-1320) %>%
  mutate(across(2:6, ~ as.numeric(trimws(.))))
head(raw_data)
```

```
## # A tibble: 6 x 6
##   Group E3 ubiquitin-protein ligase CHI~1 CCAAT/enhancer-bindi~2 'Gamma-enolase'
##   <chr>                              <dbl>                  <dbl>           <dbl>
## 1 ASD                                 619.                  1489.            733.
## 2 ASD                                 512.                  1698.           2628.
## 3 ASD                                 438.                  1122.            857.
## 4 ASD                                 505                   1210.           1394
## 5 ASD                                 441.                  1120.            885
## 6 ASD                                 499.                  1822.            658.
## # i abbreviated names: 1: 'E3 ubiquitin-protein ligase CHIP',
## #   2: 'CCAAT/enhancer-binding protein beta'
## # i 2 more variables: 'E3 SUMO-protein ligase PIAS4' <dbl>,
## #   'Interleukin-10 receptor subunit alpha' <dbl>
```

The raw data appears to be in long format consisting of 1317 proteins and 154 recorded observations for each protein. To answer the first question, we must clean the data by removing the first two rows and the second column since that information is not necessary to include right now. We want to look at the distributions of the first five proteins and compare them to the log-transformed distributions so we can also remove the the proteins not included in this sample.

```
par(mfrow = c(2, 5))
protein_sample <- raw_data[, c('E3 ubiquitin-protein ligase CHIP', 'CCAAT/enhancer-binding protein beta
abb <- c("CHIP", "CEBPB", "NSE", "PIAS4", "IL-10 Ra", "STAT3")
sum(is.na(protein_sample))
```

```
## [1] 0
```

```
# Raw values
for (i in 1:5) {
  hist(protein_sample[[i]],
       main = abb[i],
       xlab = "Protein Levels",
```

```
       col = "blue",
       border = NA,
       breaks = 30)
}
title(main = "Raw Distribution of Protein Levels", outer = TRUE, line = -1)

# Log-transformed values
log_transformed <- log10(protein_sample)
head(log_transformed)
```

```
## # A tibble: 6 x 5
##   `E3 ubiquitin-protein ligase CHIP` `CCAAT/enhancer-binding pr~1 `Gamma-enolase`
##                                <dbl>                       <dbl>           <dbl>
## 1                             2.79                        3.17            2.86
## 2                             2.71                        3.23            3.42
## 3                             2.64                        3.05            2.93
## 4                             2.70                        3.08            3.14
## 5                             2.64                        3.05            2.95
## 6                             2.70                        3.26            2.82
## # i abbreviated name: 1: `CCAAT/enhancer-binding protein beta`
## # i 2 more variables: `E3 SUMO-protein ligase PIAS4` <dbl>,
## #   `Interleukin-10 receptor subunit alpha` <dbl>
```

```
for (i in 1:5) {
  hist(log_transformed[[i]],
       main = abb[i],
       xlab = "Protein Levels",
       col = "lightgreen",
       border = NA,
       breaks = 30)
}
title(main = "Log-Transformed Protein Levels", outer = TRUE, line = -18)
```

The raw protein levels spans a wide range of values, from very low to very high concentrations. This results in a right-skewed distribution where most values are low but a few are extremely high which disproportionately affects the distribution of the data. Therefore, it makes sense to apply a log10 transformation in order to normalize the data, reduce variability, and improve model performance. After transforming, the distributions of the sample proteins are more symmetric and compressed.

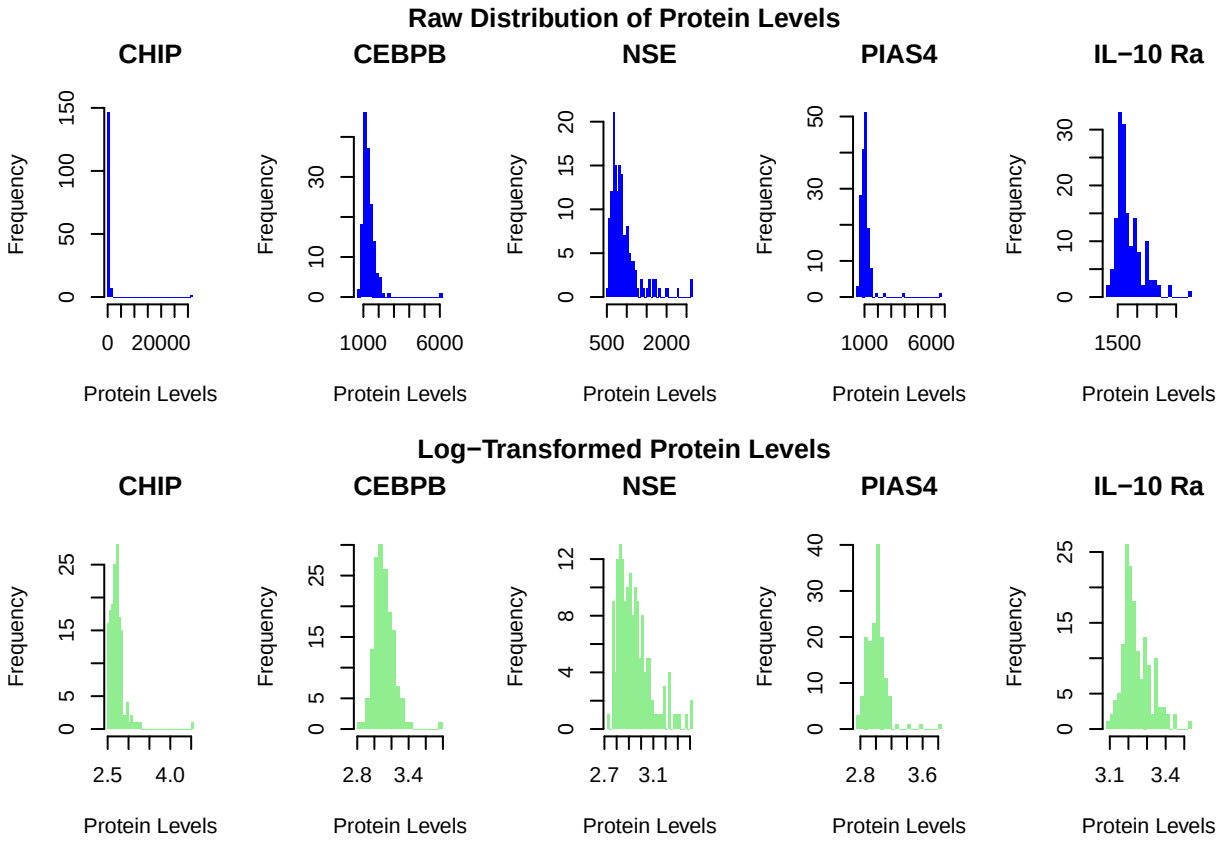

Figure 1: From left to right, the target full names are: 'E3 ubiquitin-protein ligase CHIP', 'CCAAT/enhancer-binding protein beta', 'Gamma-enolase', 'E3 SUMO-protein ligase PIAS4', and 'Interleukin-10 receptor subunit alpha'.