lucas_analysis

Lucas Childs

2025-11-03

1. What is the reason for log transforming protein levels in `biomarker-raw.csv`?

```
set.seed(1234)
library(here)
```

```
## here() starts at /Users/lucaschilds/PSTAT197A/module-1-biomarker-data-table13
```

```
rawdata <- read.csv(here("data", "biomarker-raw.csv"))

# random sample of 4 proteins to look at distributions of their levels
rand_indices <- sample(3:ncol(rawdata), 4)

prot1d <- as.numeric(rawdata[2:nrow(rawdata), rand_indices[1]])
prot1 <- prot1d[!is.na(prot1d)]

prot2d <- as.numeric(rawdata[2:nrow(rawdata), rand_indices[2]])
prot2 <- prot2d[!is.na(prot2d)]

prot3d <- as.numeric(rawdata[2:nrow(rawdata), rand_indices[3]])
prot3 <- prot3d[!is.na(prot3d)]

prot4d <- as.numeric(rawdata[2:nrow(rawdata), rand_indices[4]])
prot4 <- prot4d[!is.na(prot4d)]

summary(prot1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2568    4649    5169    5243    5668    7435
```

```
summary(prot2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3317    8790   12179   17164   18800  122168
```

```
summary(prot3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   283.1   378.7   414.2   457.8   496.9  1894.7
```

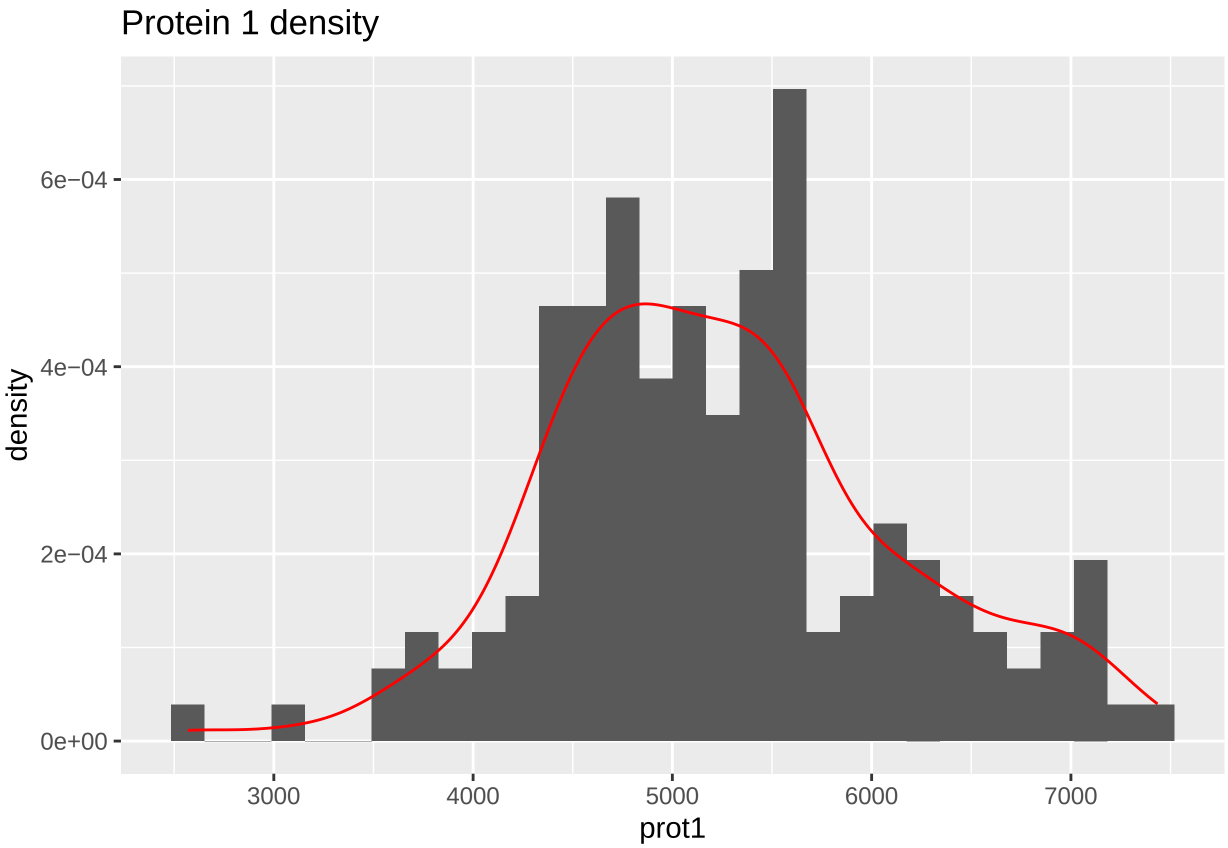
```r
summary(prot4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   954.9  5434.7  6539.2  6551.7  7719.1 10917.8
```

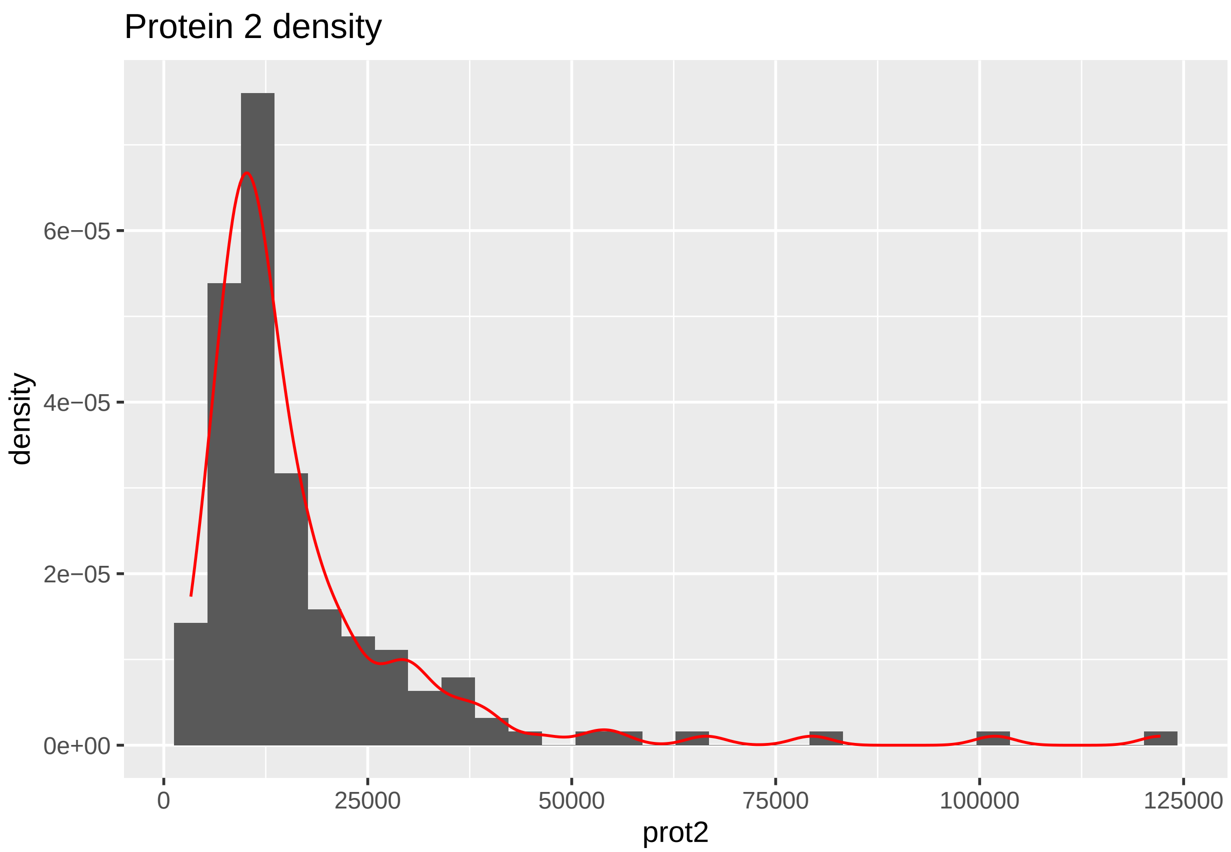
Looking at the summary statistics for 4 randomly selected proteins, the mean values differ significantly, with `prot1`'s mean being 17,164 and `prot3`'s mean being ≈ 458. `prot2`'s mean is roughly 5000 units greater than its median as well, indicating a right skew.

The log transformation of the protein levels helps compress the scale of protein levels since we have a wide range of positive values, where some are very large. Additionally, the logarithm helps with skewed data, and we have evidence that some of the data is skewed, since `prot2` has a mean much larger than its median.

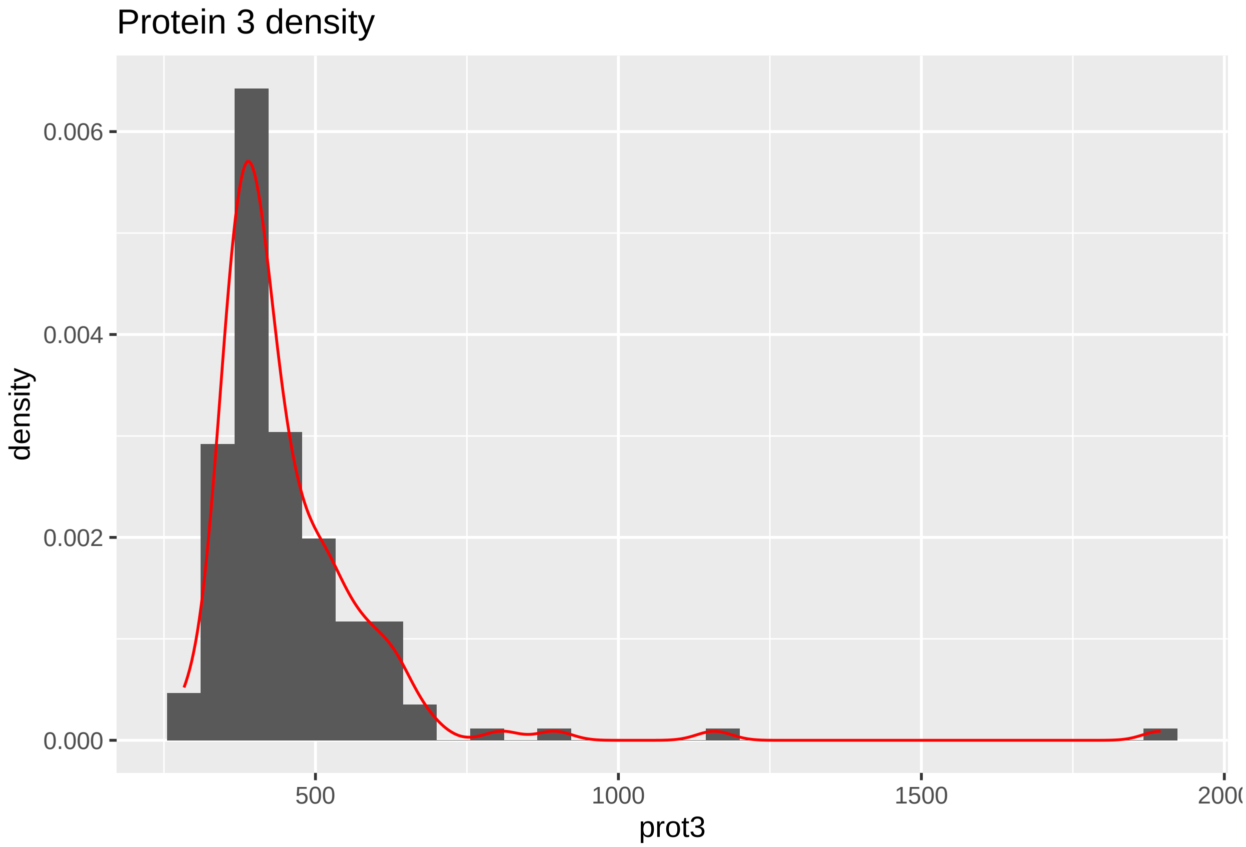
```r
library(ggplot2)
ggplot(as.data.frame(prot1), aes(x = prot1)) +
  geom_histogram(aes(y=after_stat(density)), bins = 30) +
  geom_density(color="red") +
  ggtitle('Protein 1 density')
```



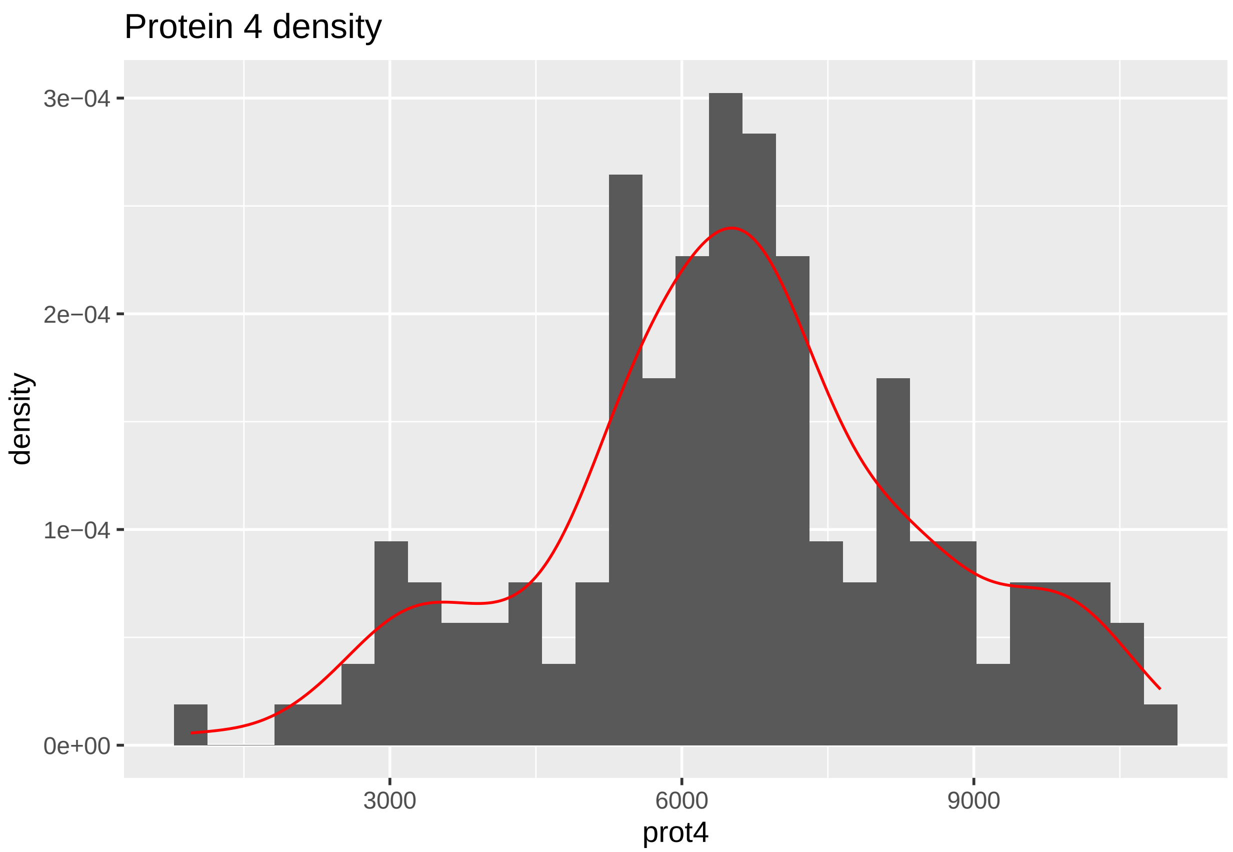
```r
ggplot(as.data.frame(prot2), aes(x = prot2)) +
  geom_histogram(aes(y=after_stat(density)), bins = 30) +
  geom_density(color="red") +
  ggtitle('Protein 2 density')
```



```
ggplot(as.data.frame(prot3), aes(x = prot3)) +
  geom_histogram(aes(y=after_stat(density)), bins = 30) +
  geom_density(color="red") +
  ggtitle('Protein 3 density')
```



```r
ggplot(as.data.frame(prot4), aes(x = prot4)) +
  geom_histogram(aes(y=after_stat(density)), bins = 30) +
  geom_density(color="red") +
  ggtitle('Protein 4 density')
```



From the density plots, `prot2` looks the most skewed (strongly right-skewed) `prot3` looks slightly right-skewed as well. Both proteins contain large outliers, however the scale of `prot2`'s protein level is much higher, so the same follows for its outliers.