question_1

Shenyi Jiang

2025-10-26

Question 1

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- 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 <- read_csv("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.
```

```
# non-protein columns
meta_cols <- c("Group", "Target Full Name")
# protein columns
protein_cols <- setdiff(names(raw)[map_lgl(raw, is.numeric)], meta_cols)
```

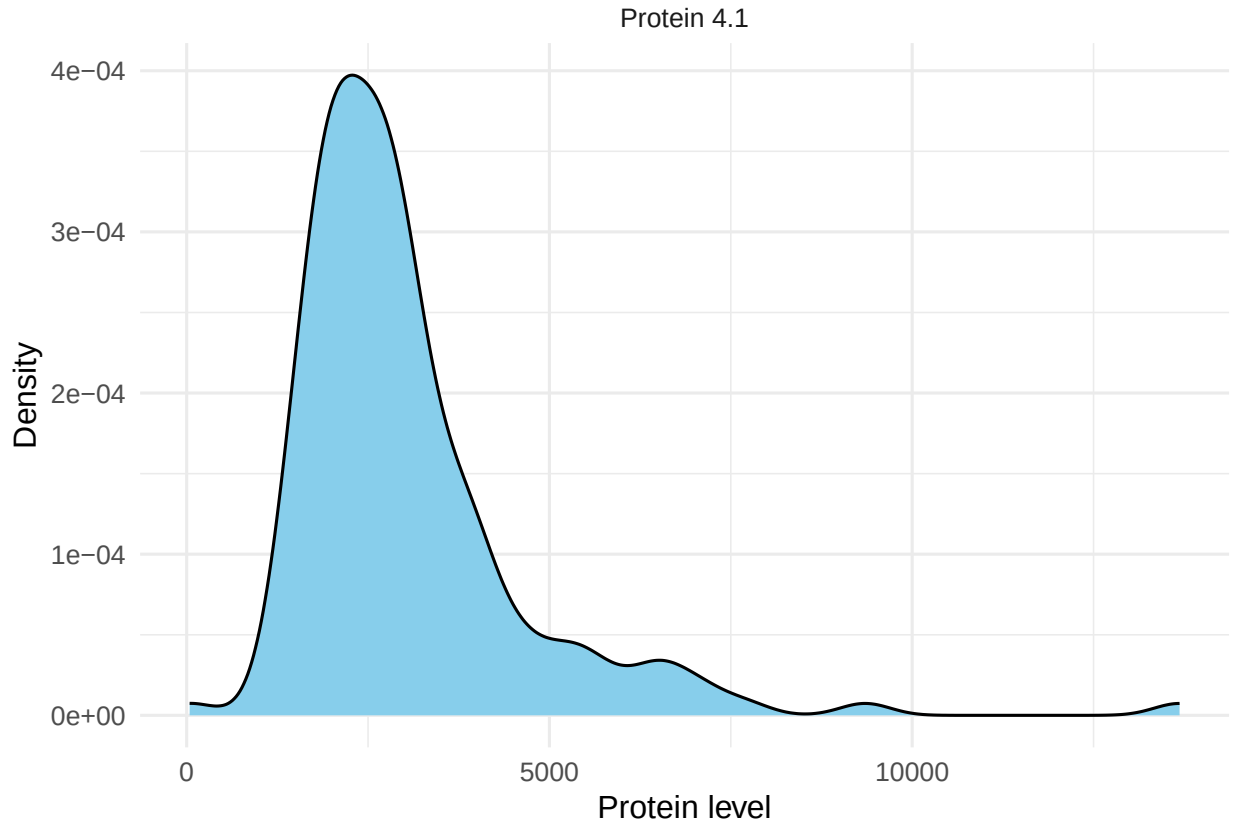
```
set.seed(1026)
sample_cols <- sample(protein_cols, min(6, length(protein_cols)))
```

```
# original data
raw %>%
  pivot_longer(all_of(sample_cols), names_to = "protein", values_to = "value") %>%
  ggplot(aes(x=value)) +
```

```
  geom_density(fill="skyblue") +
  facet_wrap(~protein, scales="free") +
  labs(titl="Raw Protein Level Distributions",
       x = "Protein level",
       y = "Density") +
  theme_minimal(base_size=12)
```

```
## Ignoring unknown labels:
## * titl : "Raw Protein Level Distributions"
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_density()`).
```




```
# log-transforming data
raw %>%
  pivot_longer(all_of(sample_cols), names_to = "protein", values_to = "value") %>%
  filter(is.finite(value)) %>%
  mutate(log_value = log1p(value),
         scaled_value = as.numeric(scale(log_value))
         ) %>%
  ggplot(aes(x=scaled_value)) +
  geom_density(fill="lightcoral") +
  facet_wrap(~protein, scales="free") +
  labs(titl="Log-transformed Protein Level Distributions",
```

```
       x = "Protein level",
       y = "Density") +
  theme_minimal(base_size=12)
```

```
## Ignoring unknown labels:
## * titl : "Log-transformed Protein Level Distributions"
```