Q2 Anna

2. Temporarily remove the outlier trimming from preprocessing and do some exploratory analysis of the outlying values. Are there specific subjects (not values) that seem to be outliers? If so, are outliers more frequent in one group or the other? (Hint: consider tabulating the number of outlying values per subject.)

Removed the outlier trimming from the preprocessing.R file and generate biomarker-clean-notrim.RData.

```
 # read in data
 # biomarker_clean <- read_csv('data/biomarker-raw.csv',
          # skip = 2,
          # col_select = -2L,
          #col_names = c('group',
          #              'empty',
          #              pull(var_names, abbreviation),
          #              'ados'),
#           na = c('-', '')) %>%
#   filter(!is.na(group)) %>%
   # log transform, center and scale, and trim

#   mutate(across(.cols = -c(group, ados),
#                 ~ scale(log10(.x))[, 1])) %>%  # no trimming

   # reorder columns
#   select(group, ados, everything())

 # export as r binary
#   save(list = 'biomarker_clean',
#      file = 'data/biomarker-clean-notrim.RData')  # new file for Q2
```

Load the new untrimmed data:

```
library(dplyr)
library(here)
library(knitr)

load(here("data", "biomarker-clean-notrim.RData"))
bm_notrim_clean <- biomarker_clean

# Define protein columns
protein <- setdiff(names(bm_notrim_clean), c("group","ados"))

# Show a small subset of the data
kable(bm_notrim_clean[1:10, c("group","ados", protein[1:8])])
```

group	ados	CHIP	CEBPB	NSE	PIAS4	IL-10 Ra	STAT3	IRF1	c-Jun
ASD	8	0.3350091	0.5203026	-0.5542975	0.6496085	-0.3575096	0.3053281	-0.4841931	0.3085327
ASD	21	-0.0714544	1.0062742	3.4723335	1.2788183	-0.1326775	1.1336984	0.2530236	0.4079033

group	ados	CHIP	CEBPB	NSE	PIAS4	IL-10 Ra	STAT3	IRF1	c-Jun
ASD	12	-0.4060154	-0.5310368	-0.0592213	1.1293856	0.5537559	-0.3339147	0.2865227	-0.8445316
ASD	20	-0.1019412	-0.2509116	1.4732606	0.0773158	-0.7046250	0.8928279	2.6073847	-0.3722943
ASD	22	-0.3952380	-0.5359999	0.0410216	-0.2989002	-0.8300687	0.8987424	1.0143168	-0.8432803
ASD	17	-0.1263865	1.2691487	-0.8923126	0.2388372	-0.3441030	0.2161441	0.2105191	0.2205713
ASD	15	0.4862308	0.7475428	-1.0868947	0.4617621	0.5703420	-0.0681688	1.0095896	1.2116105
ASD	10	-0.9903109	-1.0979811	0.2310818	-0.8849678	-0.1510567	0.0306846	-0.0346382	-0.8910910
ASD	22	-0.1083473	5.7158217	2.3159326	3.0935489	2.7577462	1.6984557	0.2089228	4.0314702
ASD	17	0.4849324	-0.2337816	-0.6973192	-0.2862883	0.0331471	1.0108135	-0.2484442	-0.2927490

To answer if there are specific subjects (not values) that seem to be outliers, we will first count the number of outliers per subject.

We defined outliers at the protein level if its value is greater than 3 standard deviations above or below the protein's mean (e.g., $|z| > 3$), since under a normal distribution about 99.7% of values lie within 3 standard deviations.

Thus, we flagged proteins that have (z score) $|z| > 3$ as outliers.

```
out_flag <- bm_notrim_clean[, protein] %>% mutate(across(everything(), ~ abs(.x) > 3))
out_counts <- tibble(
  subject = seq_len(nrow(bm_notrim_clean)),
  group   = bm_notrim_clean$group,
  n_out   = rowSums(out_flag, na.rm=TRUE)
)
```

We then counted outlier proteins for each subject and listed the top 10 subjects with the highest counts of outlier proteins.

```
top_subjects <- out_counts %>% arrange(desc(n_out)) %>% slice_head(n = 10)
kable(top_subjects)
```

subject	group	n_out
154	TD	157
108	TD	127
9	ASD	126
121	TD	122
52	ASD	121
77	TD	114
147	TD	77
24	ASD	48
100	TD	47
150	TD	47

From this table, there seems to be several subjects that have many outliers proteins.

Subject #154 had 157 outlier proteins out of the 1,125 proteins tested, so roughly 14% of them were outliers.

Also, subjects 108, 9, 121, 52, and 77 all have over 100 proteins over the threshold we set, translating to roughly 10.1% - 11.3% of the proteins in each subject being considered an outlier.

Now flagging subject-level outliers, we will use the boxplot rule (Tukey rule), which flags points greater than Q3 + (1.5 x IQR) or less than Q1 - (1.5 x IQR).

```
cutoff <- quantile(out_counts$n_out, 0.75) + 1.5 * IQR(out_counts$n_out)
subject_outliers <- subset(out_counts, n_out >= cutoff)

kable(cutoff, caption = "Outlier subject cutoff")
```

Table 3: Outlier subject cutoff

	x
75%	29.375

```
kable(subject_outliers, caption = "Subjects considered to be outliers")
```

Table 4: Subjects considered to be outliers

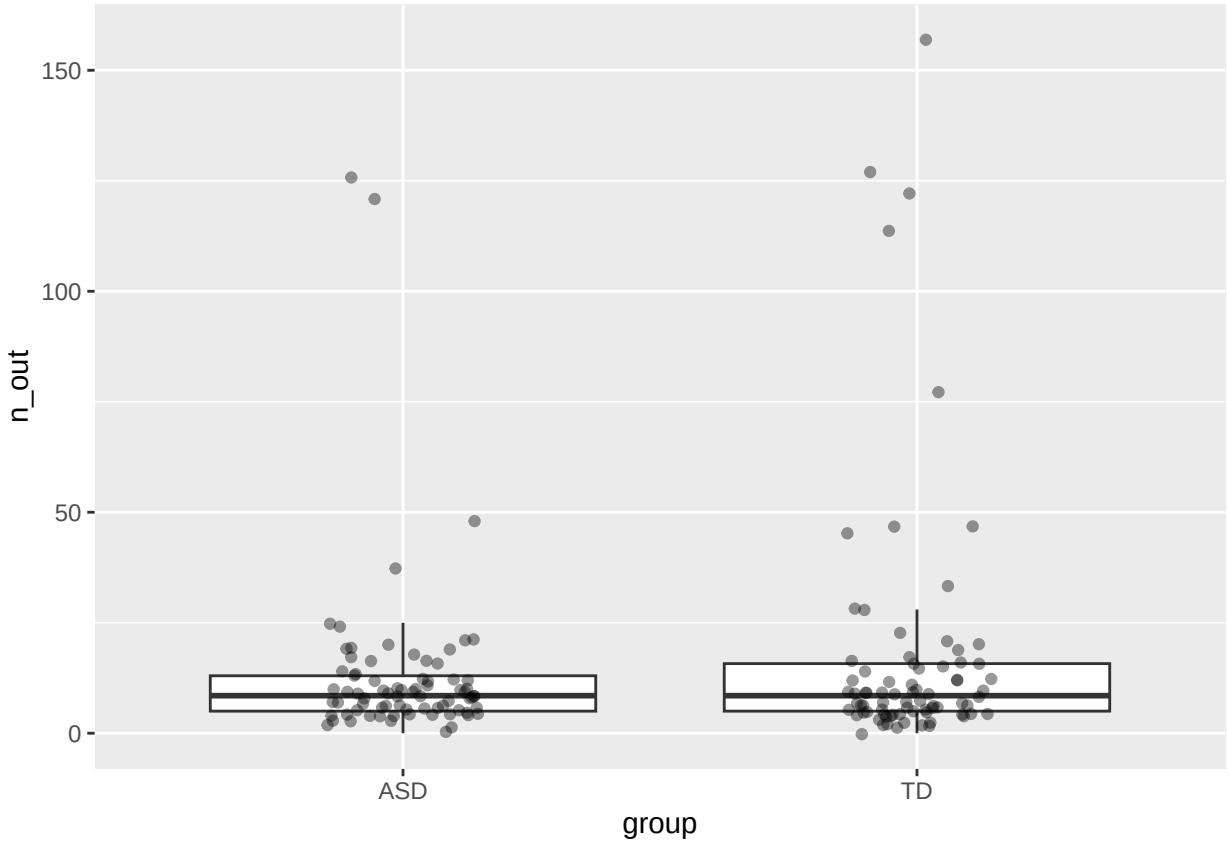
subject	group	n_out
9	ASD	126
24	ASD	48
52	ASD	121
73	ASD	37
77	TD	114
98	TD	33
100	TD	47
108	TD	127
121	TD	122
131	TD	45
147	TD	77
150	TD	47
154	TD	157

Thus, we flagged subjects with equal or more than 30 outliers proteins as subject-level outliers. They include subjects 9, 24, 52, 73, 77, 98, 100, 108, 121, 131, 147, 150, and 154.

Now, answering if outliers are more frequent in one group or the other:

Visually comparing the distributions of outlier proteins per subject by group. Each point is a subject, and the y-axis is the number of flagged outlier proteins.

```
library(ggplot2)
ggplot(out_counts, aes(group, n_out)) +
  geom_boxplot(outlier.shape=NA) + geom_jitter(width=.15, alpha=.4)
```



From the visual, it seems like the TD group has more subjects w/ higher outlier counts, including the overall maximum around 150. This supports the previous table, as 9/13 of the outlier subjects were from the TD group.

```
# Counting outlier subjects per group
x <- c(
  ASD = sum(subject_outliers$group == "ASD"),
  TD  = sum(subject_outliers$group == "TD")
)

# Total subjects per group
n <- c(
  ASD = sum(out_counts$group == "ASD"),
  TD  = sum(out_counts$group == "TD")
)

# Two-sided test to test if proportions are different
pt <- prop.test(x = x, n = n)

kable(data.frame(
    p_value = signif(pt$p.value, 4),
    prop1 = round(unname(pt$estimate[1]), 4), # prop1 = ASD
    prop2 = round(unname(pt$estimate[2]), 4) # prop2 = TD
    ),
  caption = "Two-proportion test: ASD vs TD")
```

Table 5: Two-proportion test: ASD vs TD

p_value	prop1	prop2
0.2668	0.0526	0.1154

In total, there were 76 subjects in the ASD group and 78 subjects in the TD group, so having 4 ASD outlier subjects and 9 TD outlier subjects yields 5.3% and 11.5% respectively.

However, running the two-proportion test, the p-value (0.2668) is greater than alpha = 0.05, so there's no statistically significant difference in the frequency of outlier subjects between ASD and TD.

To summarize, counting outliers per subject (values more than 3 standard deviations above or below the protein's mean) showed there are subjects with many outlying proteins (e.g., 154, 108, 9, 121, 52, 77). A quick proportion calculation found there to be a higher proportion of outlier subjects within the TD group compared to the ASD group (11.5% versus 5.3% respectively.), however, the two-proportion test found no statistically significant difference in outlier frequency.