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.)

preprocessing.R file code: outlier trimming removed

```
 # 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 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
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

EDA:

Counting the number of outliers per subject.

A data point is considered an outlier if its value is greater than 3 standard deviations above or below the protein's mean, where 99.7% of the distribution lies under a normal curve.

Thus, flagging outliers that have (z score) $|z| > 3$.

```
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)
)
```

The top 10 subjects with the highest 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

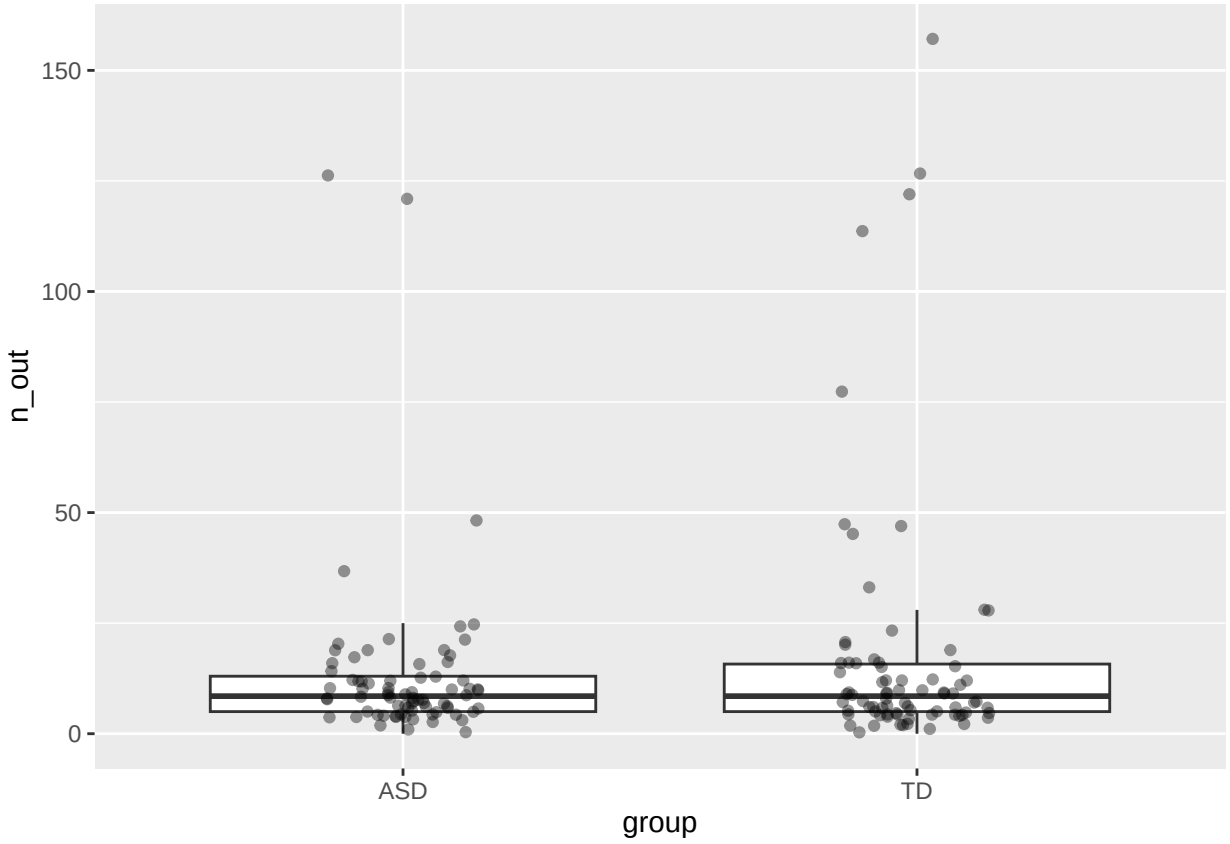
From this table, there seems to be several subjects that are outliers. Subject #154 had 157 proteins whose z-scores exceeded +/- 3. Thus, out of 1,125 proteins, roughly 14% of them were outliers.

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.

Are outliers more frequent in one group or the other? (Hint: consider tabulating the number of outlying values per subject.)

Comparing the outliers of each group:

```
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 outliers, with the highest number of outliers per subject being around 150. This supports the previous table, as 7/10 of the top 10 subjects with the highest outliers were from the TD group.

Wilcoxon rank-sum test comparing the per-subject outlier counts (n_out) between the two groups (ASD vs TD).

```
result <- wilcox.test(n_out ~ group, data = out_counts, exact=FALSE)

result_df <- data.frame(
    W = unname(result$statistic),
  p_value = signif(result$p.value, 4)

)

kable(result_df)
```

W	p_value
3028.5	0.8166

From the Wilcoxon rank-sum test, the p-value of 0.8166 > the alpha level of 0.05, so we conclude there is no evidence of a significant difference in outlier counts between the ASD and TD groups.

Conclusion:

Counting outliers per subject (outliers are defined as 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). However, a Wilcoxon rank-sum test found no evidence of a group difference in outlier counts between ASD and TD (p value = 0.8166).