EXPLORATORY DATA ANALYSIS

DSC1105

Formative Assessment 1

BILLONES, Cristel Kaye P.
2021016541

```
file_path <- "C:/Users/Cipher/Desktop/EDA_CRISTEL_FA2/cytof_one_experiment.csv"

# Read the CSV file
df <- read.csv(file_path)

# View the dataframe
View(df)
```

Data	
df	50000 obs. of 35 variables
Values	
file_path	"C:/Users/Cipher/Desktop/EDA_CRISTEL_FA2/cytof_o...

	NKp30	KIR3DL1	NKp44	KIR2DL1	GranzymeB	CXCR6	CD161	KIR2DS4	NKp46	NKG2D	NKG2C	X2B4
1	0.18759549	3.615693239	-0.560569378	-0.29366542	2.47789290	-0.14470053	-0.315287222	1.944970461	4.0818316	2.620078401	-0.35738171	-0.2711
2	1.03485177	1.700182015	-0.288961140	-0.47982795	3.26101610	-0.03392447	-0.411212891	3.802517135	3.7339299	-0.483278788	-0.46759842	-0.5594
3	2.99963976	6.141141857	1.903260569	0.48231016	4.27756173	1.94654156	-0.502234681	-0.320101715	4.5594631	-0.506908969	2.61937825	-0.4554
4	4.29985945	-0.221158600	0.242570683	-0.48312672	3.35180809	0.92622195	3.877237037	-0.169694865	4.4831486	1.927229018	-0.31101456	1.6350
5	-0.43864477	-0.503589202	-0.152632039	0.75061281	3.19414532	-0.05893640	1.090737874	-0.050330253	0.8379358	-0.458167375	0.92169475	1.2419
6	2.08830498	-0.399264593	3.455067648	-0.52008557	4.34510247	-0.36434277	-0.570589119	-0.450335915	4.0550848	3.428356456	0.62728370	-0.4157
7	-0.61325960	-0.116638004	-0.451210998	3.54585152	1.54305965	-0.41351724	0.725491115	-0.067359586	2.6651401	-0.083767345	-0.40203583	0.4703
8	-0.34138934	-0.253412447	-0.459217322	2.89927136	-0.54519329	-0.61175860	-0.127954802	2.791525907	0.7255918	2.095569158	-0.06029957	2.7497
9	2.31165616	-0.364033566	-0.572780674	3.83522214	2.92090697	2.45722962	2.247926349	-0.376103921	4.3333043	-0.016765446	-0.22356633	-0.3901
10	3.48445371	-0.028223629	-0.182148182	4.19825757	4.73955776	0.69468032	5.083731143	4.189183791	4.5397170	2.163185593	-0.59332380	1.7286
11	4.45990965	1.863340430	-0.277722452	-0.54860750	1.57021226	-0.61953852	-0.172082885	4.773904098	3.2967321	0.053960367	-0.50775778	-0.4482

 # 1

```
library(tidyr)

df$cell_id <- 1:nrow(df)

reshaped_df <- pivot_longer(df,
                            cols = -cell_id,
                            names_to = "protein_identity",
                            values_to = "amount")

dim(reshaped_df)
```

```
reshaped_df        1750000 obs. of 3 variables
    $ cell_id          : int [1:1750000] 1 1 1 1 1 1 1 1 1 1 ...
    $ protein_identity: chr [1:1750000] "NKp30" "KIR3DL1" "NKp44" "K...
    $ amount           : num [1:1750000] 0.188 3.616 -0.561 -0.294 2....
Values
```

Filter

	cell_id	protein_identity	amount
1	1	NKp30	0.187595490
2	1	KIR3DL1	3.615693239
3	1	NKp44	-0.560569378
4	1	KIR2DL1	-0.293665422
5	1	GranzymeB	2.477892899
6	1	CXCR6	-0.144700528
7	1	CD161	-0.315287222
8	1	KIR2DS4	1.944970461
9	1	NKp46	4.081831583
10	1	NKG2D	2.620078401
11	1	NKG2C	-0.357381711
12	1	X2B4	-0.271155672

Showing 1 to 13 of 1,750,000 entries, 3 total columns

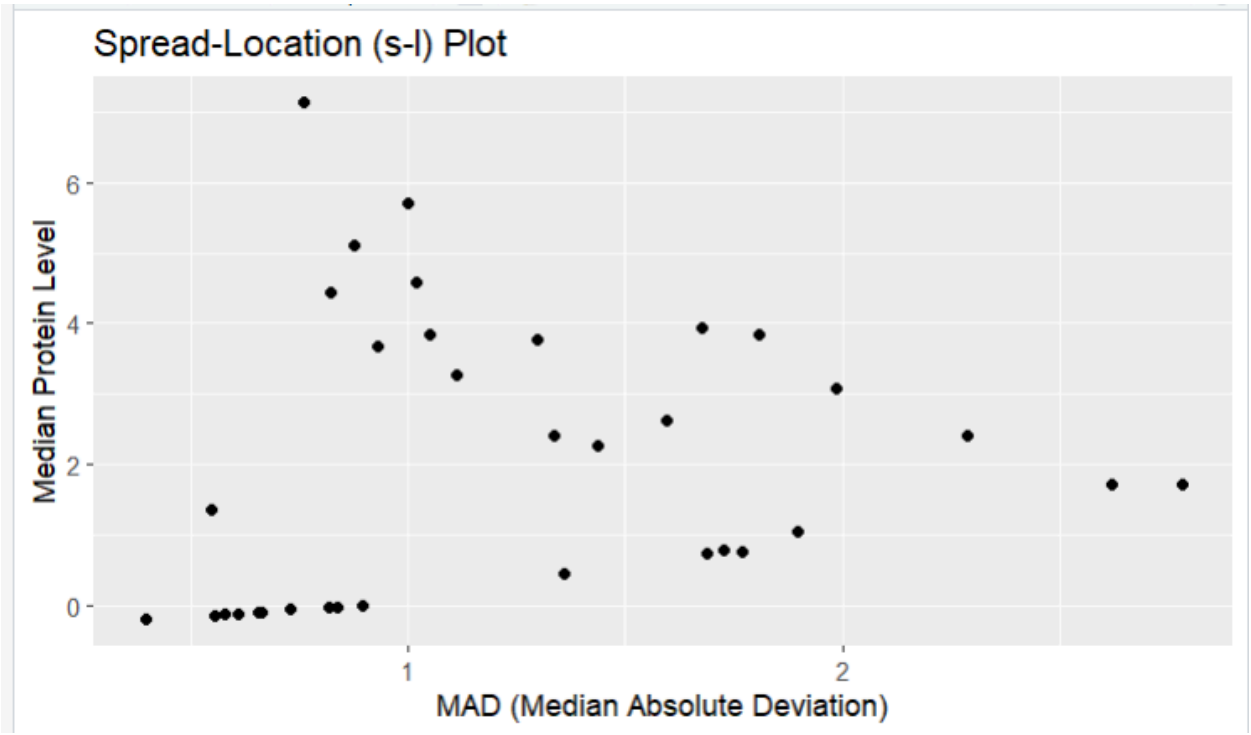
#2

```
library(dplyr)

summary_stats <- reshaped_df %>%
  group_by(protein_identity) %>%
  summarise(
    median_protein_level = median(amount),
    median_abs_dev = mad(amount)
  )

print(summary_stats)
```

```
> print(summary_stats)
# A tibble: 35 × 3
   protein_identity median_protein_level median_abs_dev
   <chr>                           <dbl>          <dbl>
 1 CD107a                         -0.122          0.609
 2 CD16                            5.12           0.874
 3 CD161                           0.726          1.69
 4 CD2                             3.95           1.68
 5 CD4                            -0.204          0.395
 6 CD56                            5.71           0.998
 7 CD57                            3.07           1.99
 8 CD69                            4.59           1.02
 9 CD8                             2.40           2.29
10 CXCR6                          -0.0581         0.727
# ℹ 25 more rows
# ℹ Use `print(n = ...)` to see more rows
>
```

#3

```
library(ggplot2)

ggplot(summary_stats, aes(x = median_abs_dev, y = median_protein_level)) +
  geom_point() +
  labs(x = "MAD (Median Absolute Deviation)",
       y = "Median Protein Level",
       title = "Spread-Location (s-l) Plot")
```



#4

```
library(tidyr)
library(dplyr)
file_path <- "C:/Users/Cipher/Desktop/EDA_CRISTEL_FA2/gymData.csv"
gymData <- read.csv(file_path)
View(gymData)
```

	country	vault_2012	floor_2012	vault_2016	floor_2016
1	United States	48.132	45.366	46.866	45.999
2	Russia	46.366	41.599	45.733	42.032
3	China	44.266	40.833	44.332	42.066

```
gymData_reshaped <- gymData %>%


  pivot_longer(cols = -country,
               names_to = c("Event", "Year"),
               names_pattern = "(.*)_(\\d{4})",
               values_to = "Score") %>%
  # Separate the Year column to only keep the year
  mutate(Year = as.integer(Year)) # convert year to integer

View(gymData_reshaped)
```

Filter

	country	Event	Year	Score
1	United States	vault	2012	48.132
2	United States	floor	2012	45.366
3	United States	vault	2016	46.866
4	United States	floor	2016	45.999
5	Russia	vault	2012	46.366
6	Russia	floor	2012	41.599
7	Russia	vault	2016	45.733
8	Russia	floor	2016	42.032
9	China	vault	2012	44.266
10	China	floor	2012	40.833
11	China	vault	2016	44.332
12	China	floor	2016	42.066