EXPLORATORY DATA ANALYSIS

DSC1105

Formative Assessment 1

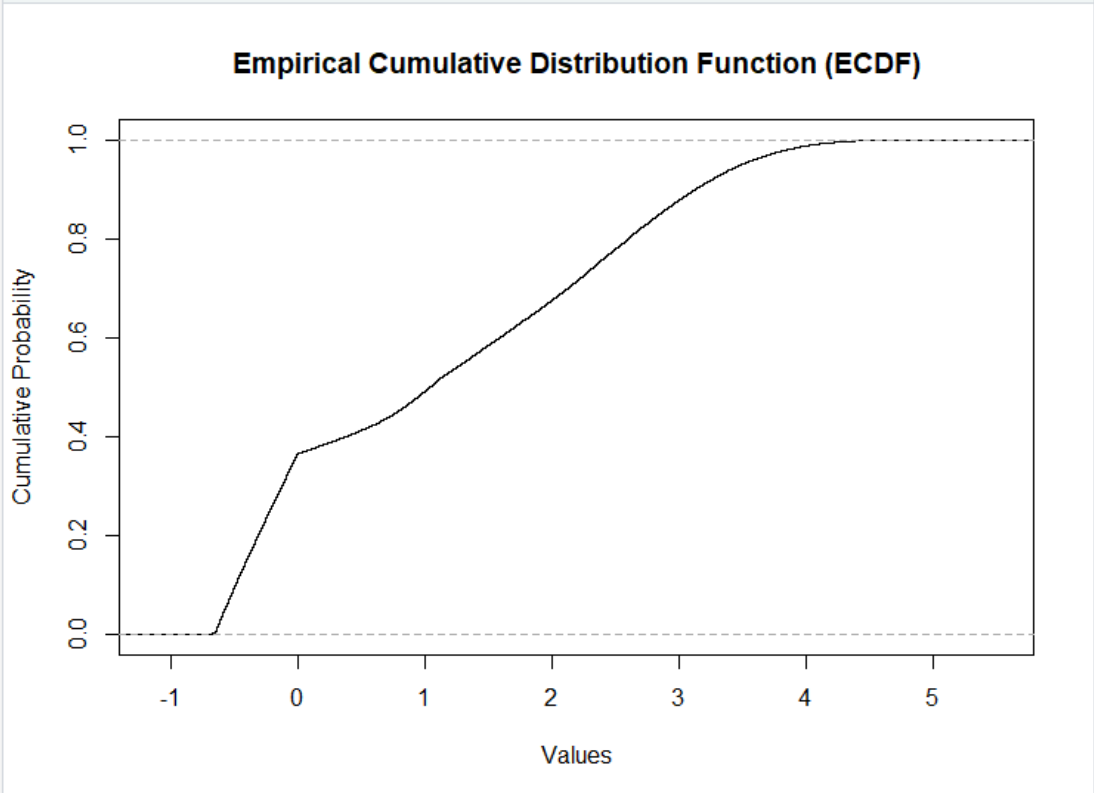
BILLONES, Cristel Kaye P.
2021016541

I.

```
cytof_one_experiment <- read.csv("C:/Users/Cipher/Desktop/CRISTEL/cytof_one_experiment.csv")

X2_col <- cytof_one_experiment$X2B4
```

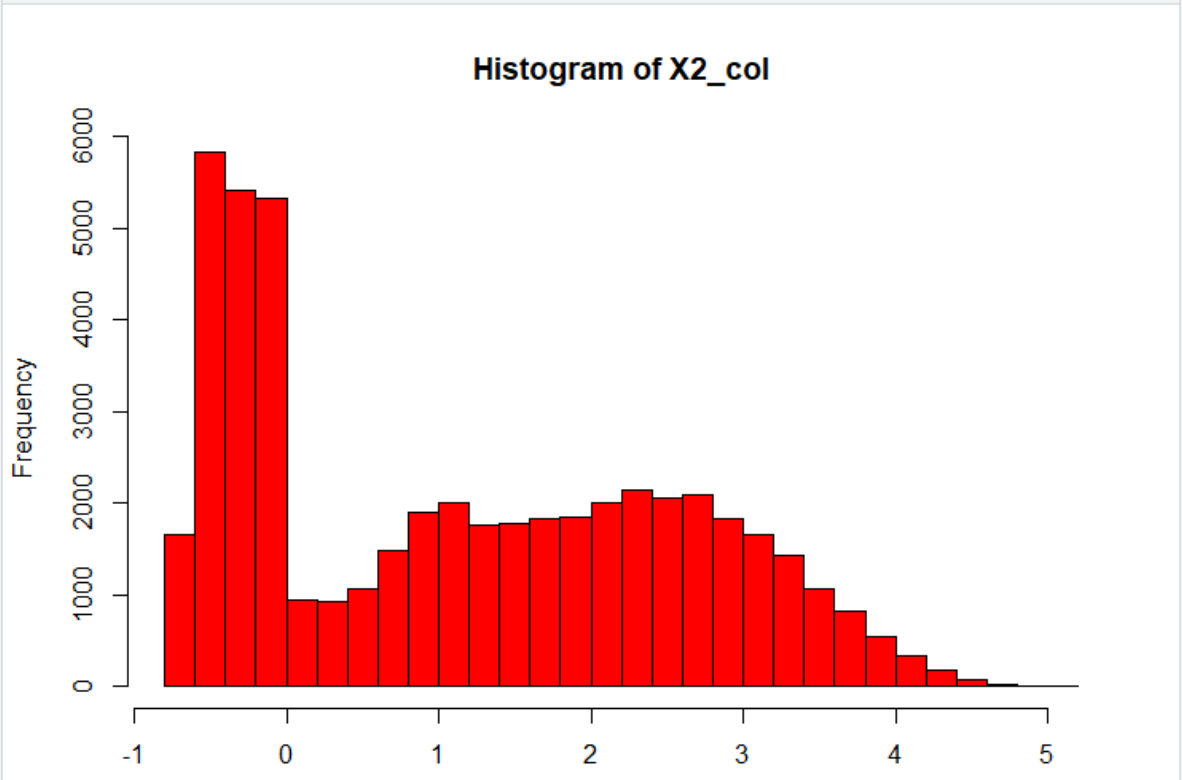
```
ecdf_values <- ecdf(X2_col)
plot(ecdf_values,
     main = "Empirical Cumulative Distribution Function (ECDF)",
     xlab = "Values",
     ylab = "Cumulative Probability"
)
```



Interpretation:

The ECDF plot pattern likely represents a positively skewed distribution, with most data points concentrated towards the lower end (left-skewed) and fewer points spread out towards the higher end (right-skewed). In conclusion, the ECDF plot reveals the distributional characteristics of your dataset, highlighting the transition from a concentrated range of lower values to a more spread-out distribution as values increase.

```r
hist(X2_col,
     breaks = "FD",
     col = "red",
     main = "Histogram of X2_col",
     xlab = "Values",
     ylab = "Frequency"
)
```



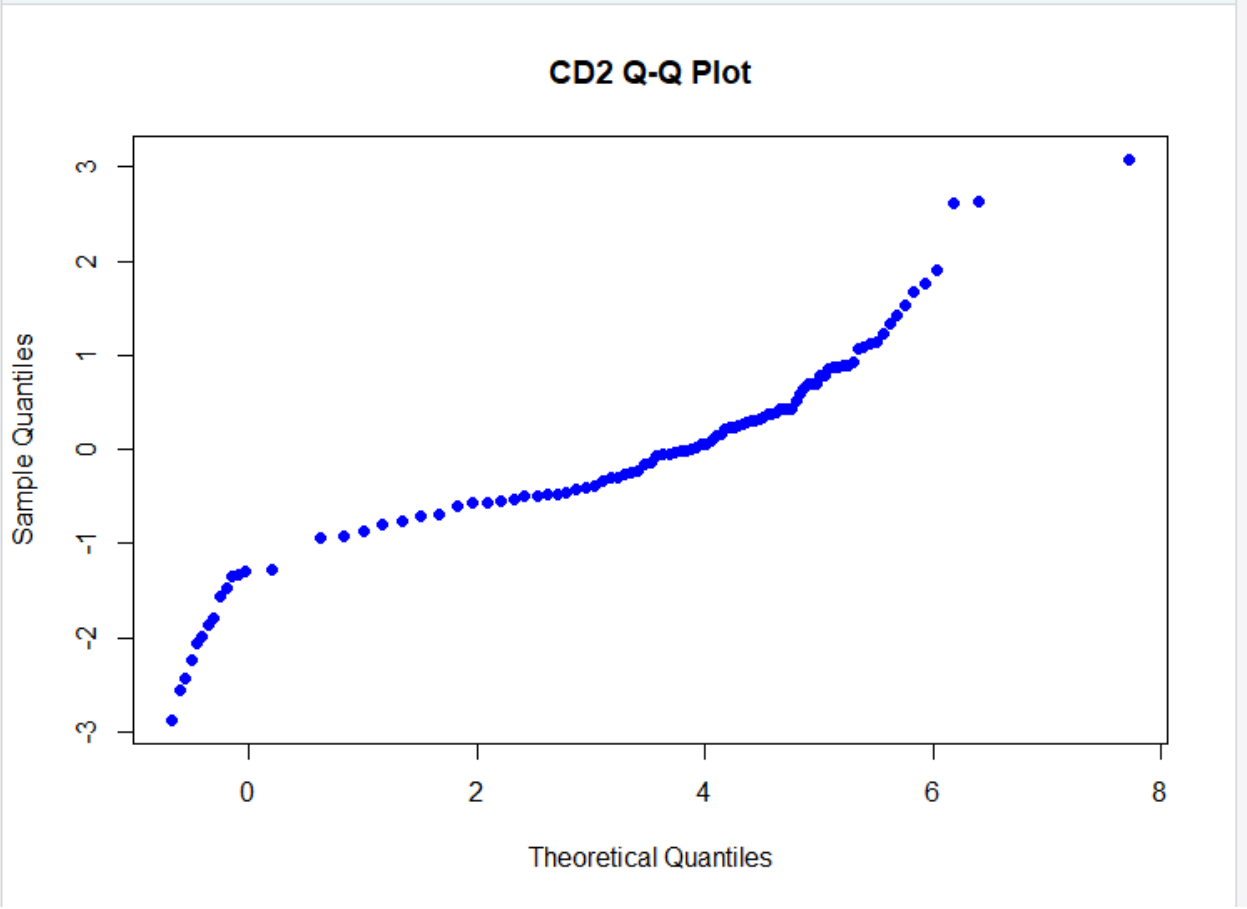
Interpretation:

The histogram's shape suggests that the dataset is skewed, likely positively skewed. This means that the majority of the data points are concentrated towards the lower end (below zero), with fewer data points spread out towards the higher end (above zero).

II.

```
CD2 <- cytof_one_experiment$CD2
CD4 <- cytof_one_experiment$CD4
```

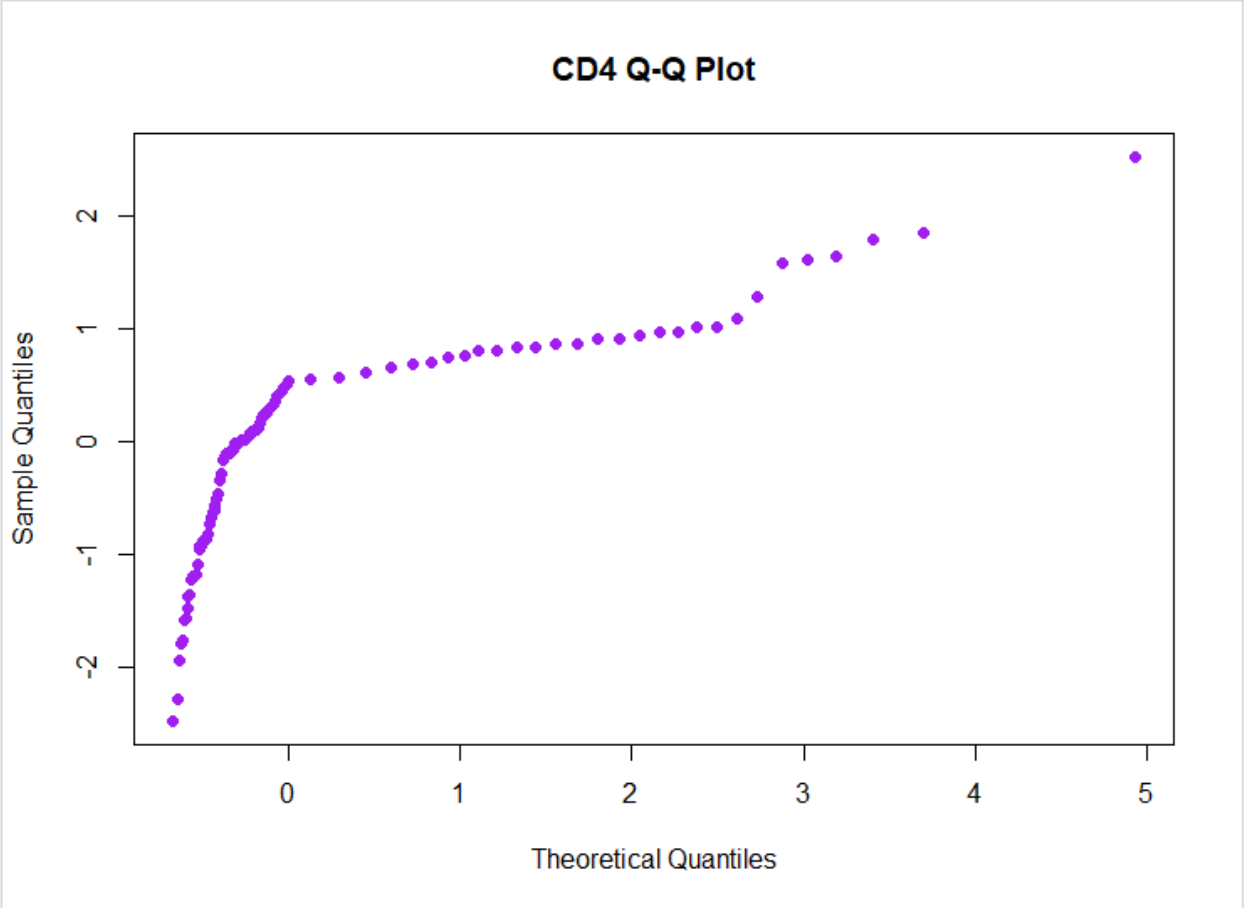
```
qqplot(CD2, rnorm(100),
       main = "CD2 Q-Q Plot",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles",
       col = "blue",
       pch = 19)
```



Interpretation:

The smooth upward slope in the Q-Q plot suggests that the first dataset is relatively normally distributed or closely follows a normal distribution.

```r
qqplot(CD4, rnorm(100),
       main = "CD4 Q-Q Plot",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles",
       col = "purple",
       pch = 19)
```



Interpretation:

The steep slope below zero in the Q-Q plot of the second dataset indicates a deviation from the normal distribution, particularly for values below zero. The steep slope implies that there might be outliers or extreme values in the second dataset that are pulling the distribution away from the theoretical normal distribution.

Summary:

The first dataset likely has minimal skewness or is symmetric, given its smooth upward slope in the Q-Q plot and its close adherence to the theoretical normal distribution. The second dataset may exhibit some degree of skewness, especially considering the steep slope below zero in its Q-Q plot, which suggests asymmetry or non-normality in its distribution.