# FA1_EDA

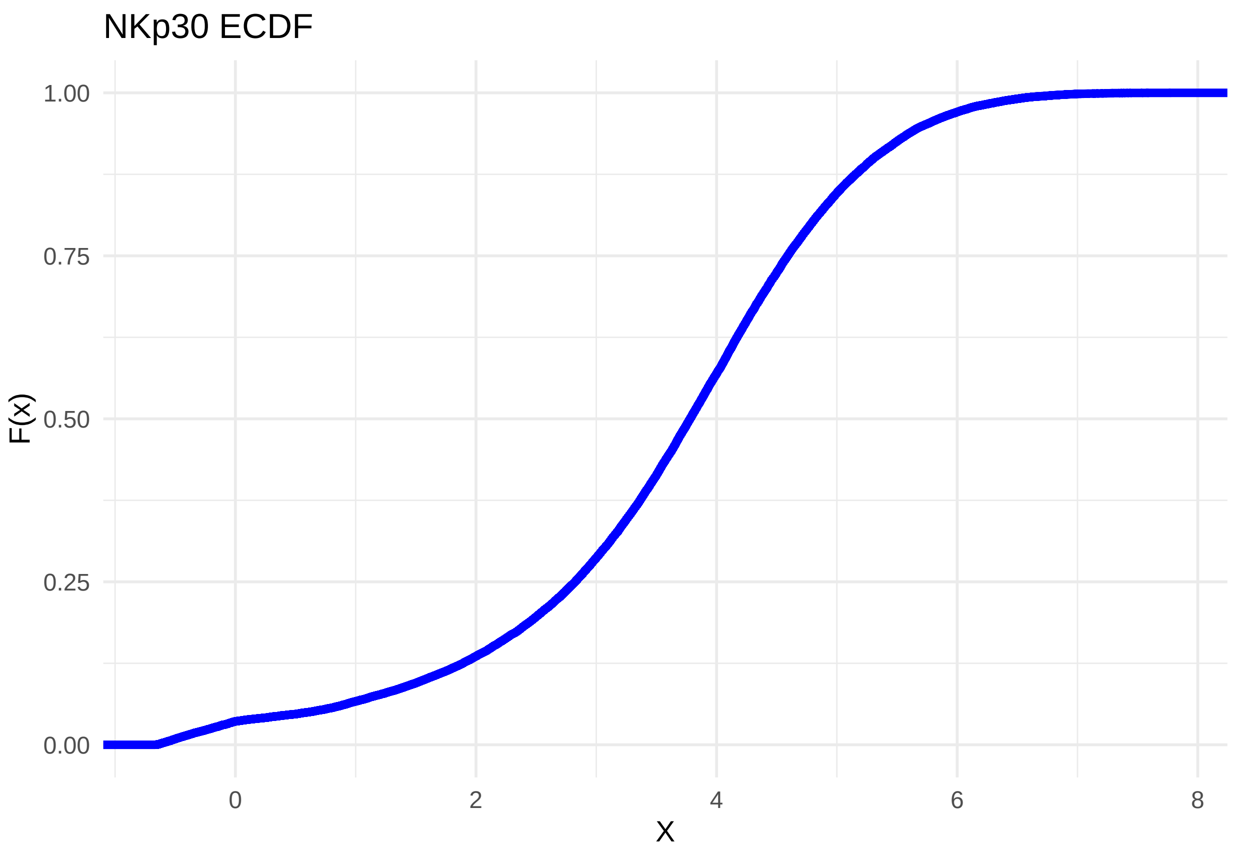
Julian Philip S. Quijano

2026-01-21

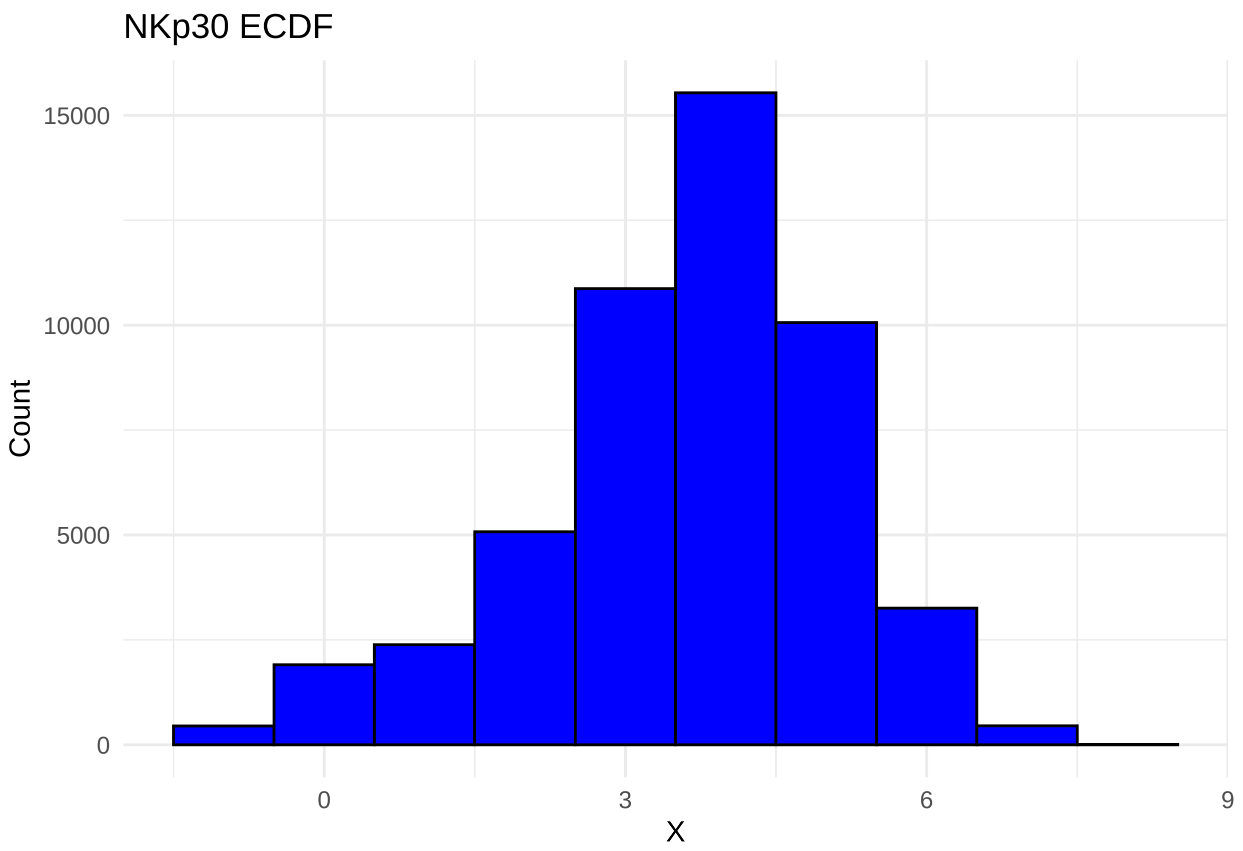
```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.3

## Warning: package 'ggplot2' was built under R version 4.4.3

## Warning: package 'tibble' was built under R version 4.4.3

## Warning: package 'tidyr' was built under R version 4.4.3

## Warning: package 'readr' was built under R version 4.4.3

## Warning: package 'purrr' was built under R version 4.4.3

## Warning: package 'dplyr' was built under R version 4.4.3

## Warning: package 'forcats' was built under R version 4.4.3

## Warning: package 'lubridate' was built under R version 4.4.3

## -- 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   1.5.1
## v ggplot2   4.0.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- 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
```

```r
#1
df <- read.csv("cytof_one_experiment.csv")

a <- ecdf(df$NKp30)

ggplot(df, aes(x = NKp30)) +
  stat_ecdf(geom = 'step', color = 'blue', linewidth = 1.5) +
  labs(title = "NKp30 ECDF", x = "X", y = 'F(x)') +
  theme_minimal()
```



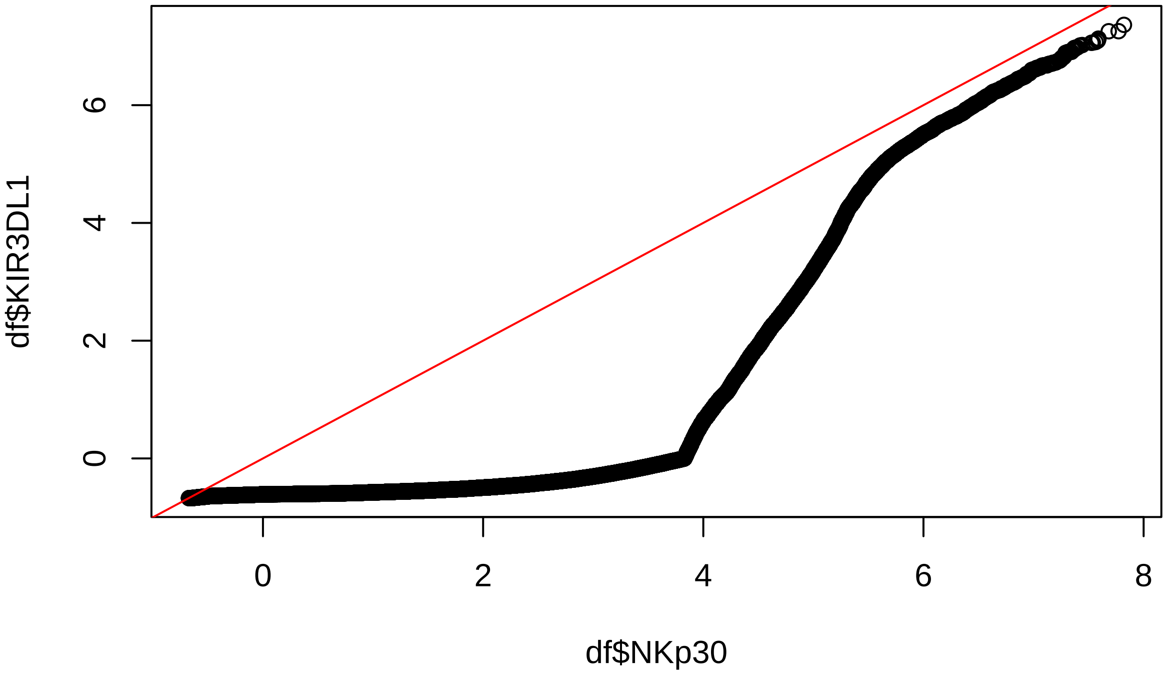
```r
ggplot(df, aes(x = NKp30)) +
  geom_histogram(binwidth = 1, fill = 'blue', color = 'black') +
  labs(title = "NKp30 ECDF", x = "X", y = 'Count') +
  theme_minimal()
```



**What does the plot tell you about the distribution of the values in that columns?**

It seems that 50% of all values in NKp30 are $\leq 3.78$, also from visual inspection, with the S shape of the ECDF and the bell shape of the histogram, NKp30 is normal and the mean falls at 3.78 and majority of the datapoints are below the mean.

```
#2

qqplot(df$NKp30, df$KIR3DL1)
abline(0, 1, col='red')
```



### What does the Q-Q plot tell you about similarities or differences between the distributions of the values in the two columns?

With the graph plateauing at the start, it shows that values in NKp30 are similar or clustered, while values in KIR3DL1 are more spread out. With the graph rising in the middle, it shows that the values from NKp30 are spreading out. Finally, the end of the line follows and almost parallels the 45 degree line, showing that the data from the two datasets have similar values by the end.

Github Link: https://github.com/SylTana/DSC1105-Exploratory-Data-Analysis/tree/main/FA1