Formative Assessment 1

DSC1105 Exploratory Data Analysis

Sindayen, Lorenzo Danilo V.

2026-01-25

```r
knitr::opts_chunk$set(
  echo = TRUE,
  warning = FALSE,
  message = FALSE
)
```

```r
library(tidyverse)
library(knitr)
library(ggplot2)
```

GitHub Link:
https://github.com/ChewyGnome/APM1111

Dataset

```r
cytof <- read_csv("cytof_one_experiment (1).csv", show_col_types = FALSE)
```

```r
kable(
  head(cytof[, 1:6]),
  caption = "First Six Observations of Selected CyTOF Markers"
)
```

Table 1: First Six Observations of Selected CyTOF Markers

NKp30	KIR3DL1	NKp44	KIR2DL1	GranzymeB	CXCR6
0.1875955	3.6156932	-0.5605694	-0.2936654	2.477893	-0.1447005
1.0348518	1.7001820	-0.2889611	-0.4798280	3.261016	-0.0339245
2.9996398	6.1411419	1.9032606	0.4823102	4.277562	1.9465416
4.2998594	-0.2211586	0.2425707	-0.4831267	3.351808	0.9262220
-0.4386448	-0.5035892	-0.1526320	0.7506128	3.194145	-0.0589364

NKp30	KIR3DL1	NKp44	KIR2DL1	GranzymeB	CXCR6
2.0883050	-0.3992646	3.4550676	-0.5200856	4.345103	-0.3643428

Introduction

Exploratory Data Analysis is a method for understanding the structure and characteristics of data through the method of graphical and numerical summaries. It is an analycal tool used primarily for revealing patterns, distributional features, and potential irregularities without imposing strong modeling assumptions.

In this assessment, the researcher explores the distribution of a single CyTOF marker,comparing the distributions of two markers using visualization techniques discussed in lecture.

Problem(s)

1. Examine the distribution of one CyTOF marker using at least two univariate plots
2. What does the plot tell you about the distribution of the values in that column?
3. What does the Q-Q plot tell you about similarities or differences between the distributions of the values in the two columns?

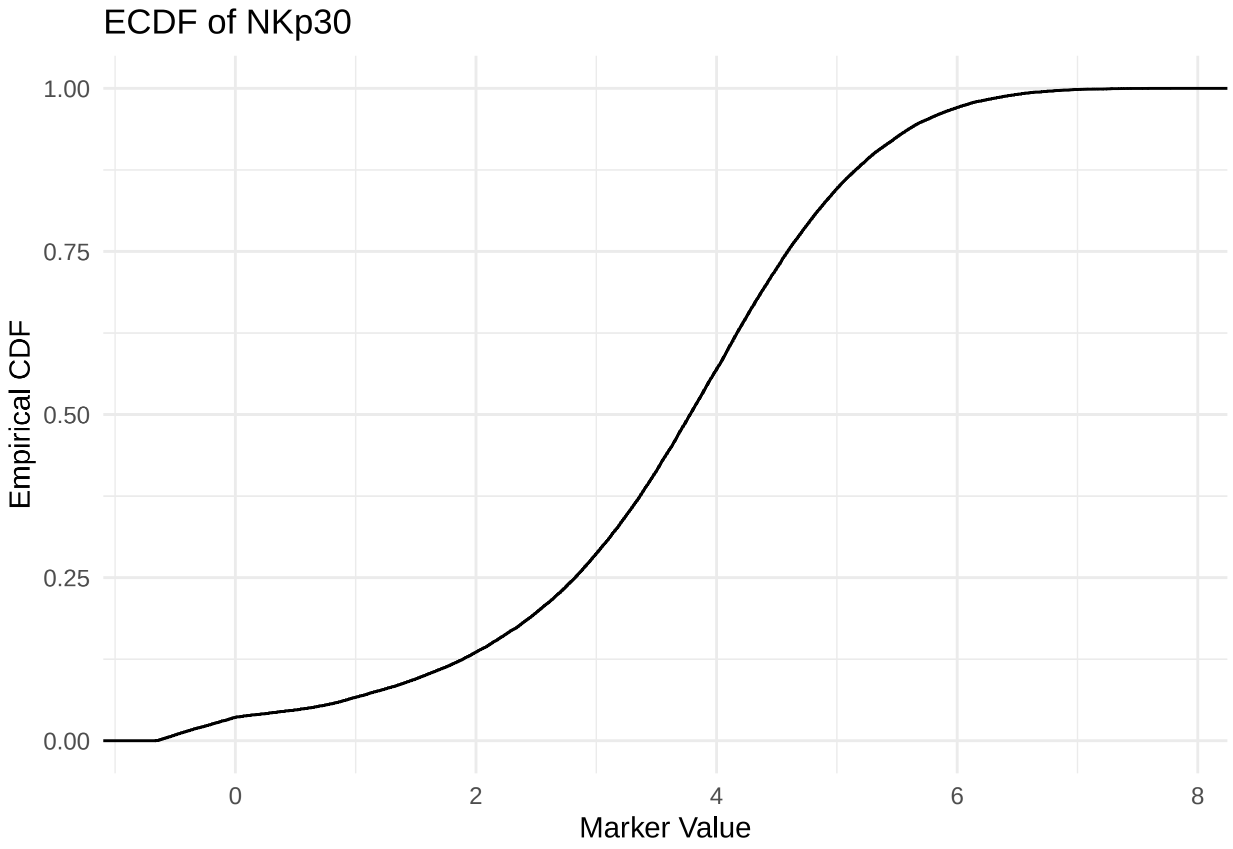
Calculations of Data

Univariate Analysis

```r
first_col_name <- colnames(cytof)[1]
first_col <- cytof[[1]]
```

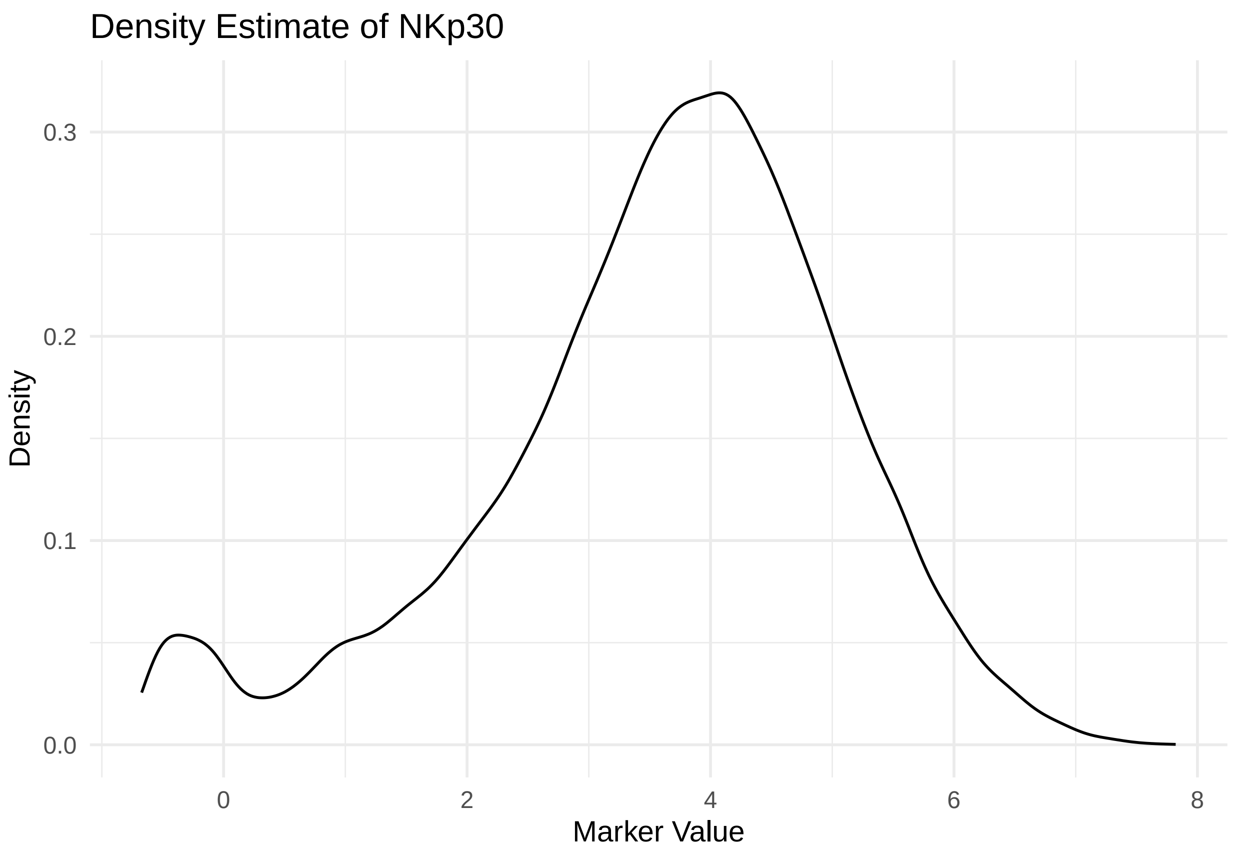
Empirical Cumulative Distribution Function

```r
ggplot(cytof, aes(x = .data[[first_col_name]])) +
  stat_ecdf() +
  labs(
    title = paste("ECDF of", first_col_name),
    x = "Marker Value",
    y = "Empirical CDF"
  ) +
  theme_minimal()
```



Density Estimate

```r
ggplot(cytof, aes(x = .data[[first_col_name]])) +
  geom_density() +
  labs(
    title = paste("Density Estimate of", first_col_name),
    x = "Marker Value",
    y = "Density"
  ) +
  theme_minimal()
```

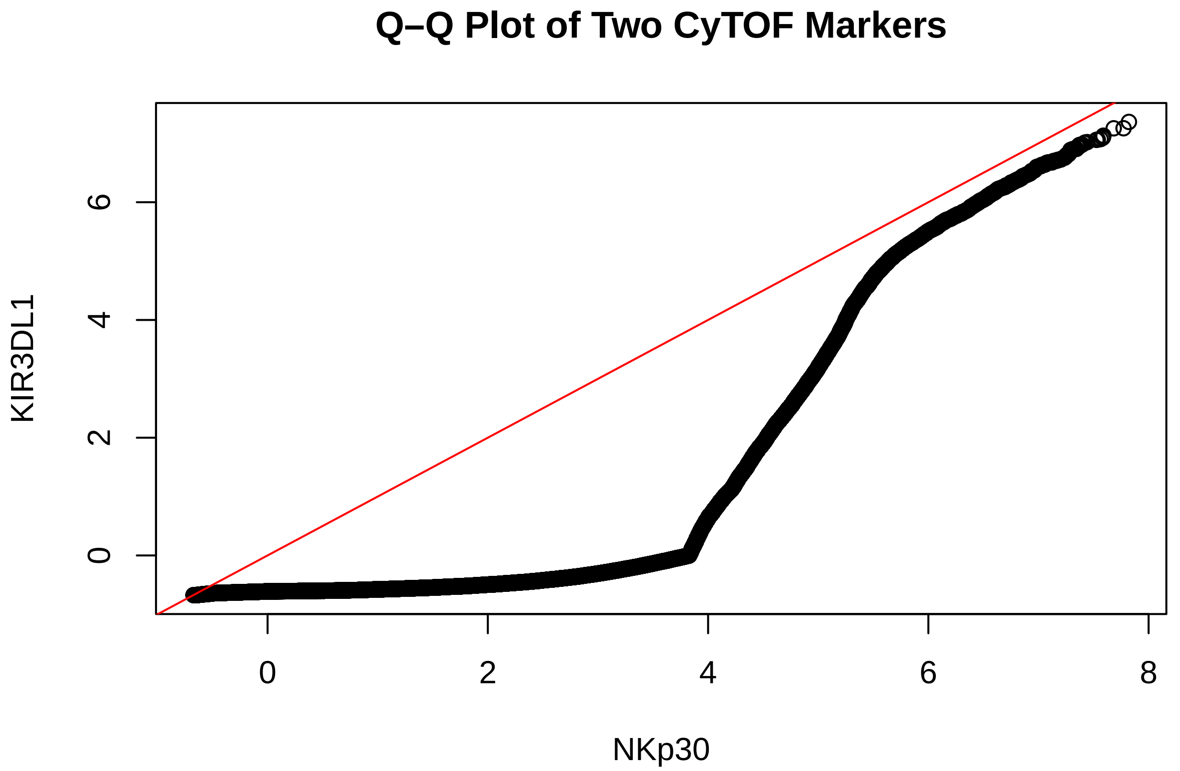


Q–Q Plot Comparison

For the comparison, the first and the second columns were selected from the CyTOF dataset

```
second_col_name <- colnames(cytof)[2]
second_col <- cytof[[2]]
```

```
qqplot(first_col, second_col,
       xlab = first_col_name,
       ylab = second_col_name,
       main = "Q-Q Plot of Two CyTOF Markers")
abline(0, 1, col = "red")
```



Conclusion

In this study, the researcher employed the exploratory data analysis technique as a way to assess the selected CyTOF columns. In this test univariate visualizations were used to summarize the behavior of a single marker, on the other hand the Q–Q plot showed the comparison between the distributions of two markers. Overall, the results show a high importance of visualization when working with high-dimensional biological data such as CyTOF measurements.