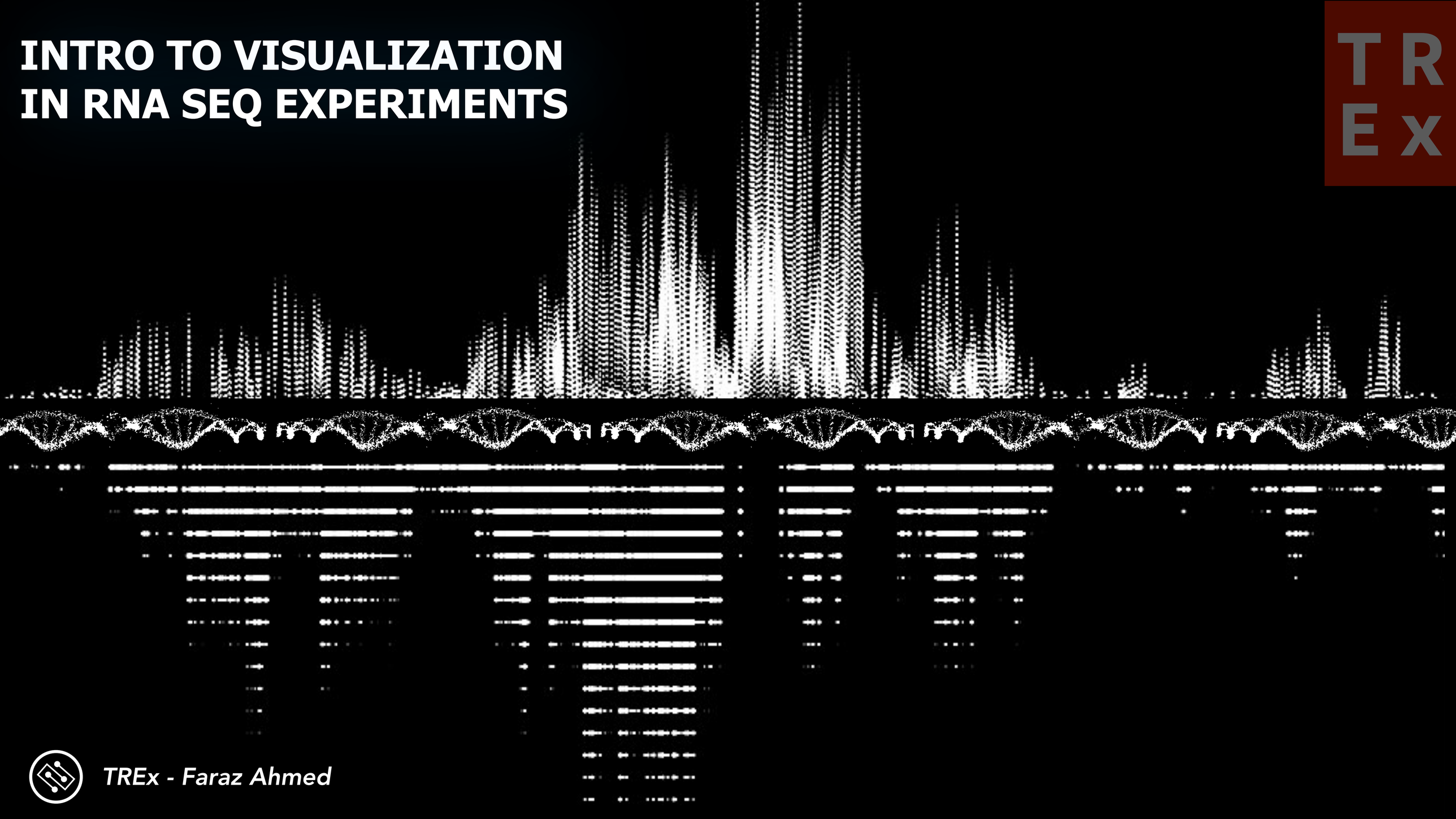
# INTRO TO VISUALIZATION IN RNA SEQ EXPERIMENTS



TREx

*TREx - Faraz Ahmed*

# A WORLD OF POSSIBILITIES

Violin

Density

Histogram

Boxplot

Ridgeline

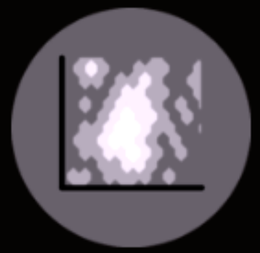
Scatter

Heatmap

Correlogram

Bubble

Connected scatter



Density 2d

Barplot

Spider / Radar


Wordcloud

Parallel

Lollipop

Circular Barplot

Treemap

Venn diagram

Doughnut

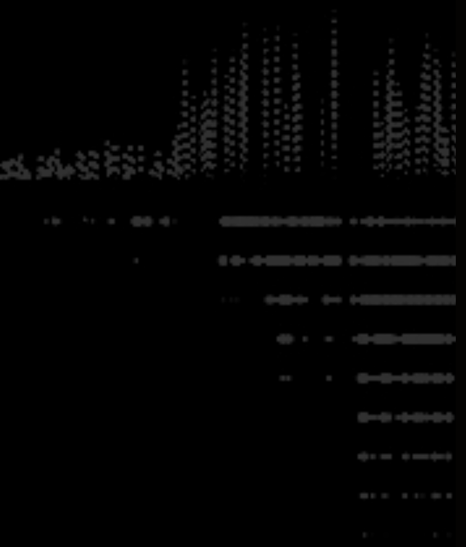
Pie chart

Dendrogram

Circular packing

Sunburst

# A WORLD OF POSSIBILITIES



Venn diagram

Doughnut

Pie chart

Dendrogram

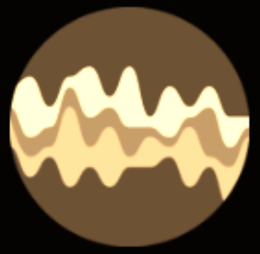
Circular packing

Sunburst

Line plot

Area

Stacked area

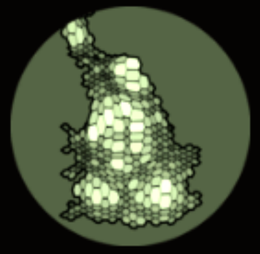


Streamchart

Map

Choropleth

Hexbin map



Cartogram

Connection

Bubble map

Chord diagram

Network

Sankey

Arc diagram

Edge bundling

# WHAT IS MY GOAL?

## WHICH *PLATFORM* DO I USE TO GENERATE MY PLOTS 🤔

- DATA EXPLORATION?
- DATA SUMMARY?

# WHICH *PLATFORM* DO I USE TO GENERATE MY PLOTS 🤔

Command Line

OR

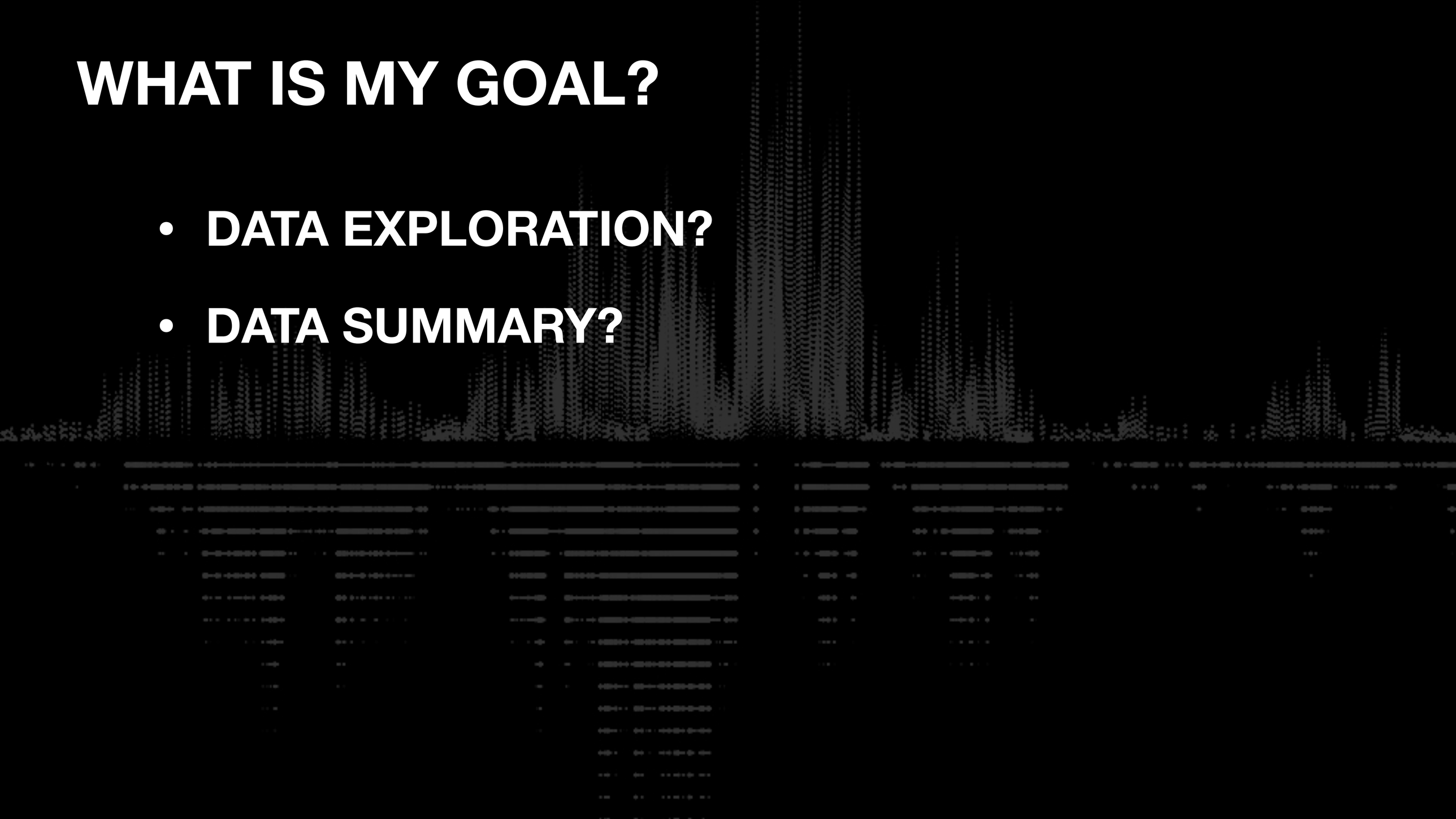
Graphical User Interface (GUI)

# WHICH *PLATFORM* DO I USE TO GENERATE MY PLOTS 🤔

- Command Line:
  - R/RSTUDIO (ggplot2, reshape, plotly, viridis)
  - Python (matplotlib, plotly)
  - jQuery
- GUI's:
  - JMP Pro
  - GraphPad Prism
  - R-shiny

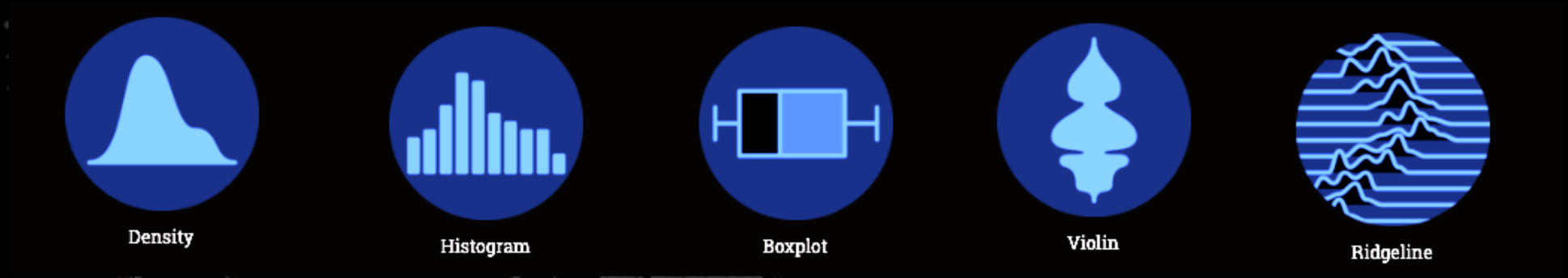
# WHAT IS MY GOAL?

- DATA EXPLORATION?
- DATA SUMMARY?



# DATA EXPLORATION

- GENERALLY MEANS, PLOTTING ALL DATA POINTS;
- FINDING PATTERNS;



# REVIEW

***DATA*** **can either be *DISCRETE* or *CONTINUOUS***

***Discrete data:***

**Can only take particular values**

**Each value is distinct (up to ∞) - NO Grey Area**

**can be numeric -- like numbers of DE genes**
**but it can also be categorical -- like case or control,**
**or male or female, or WT or KO.**

# REVIEW

***DATA*** **can either be** ***DISCRETE*** **or** ***CONTINUOUS***

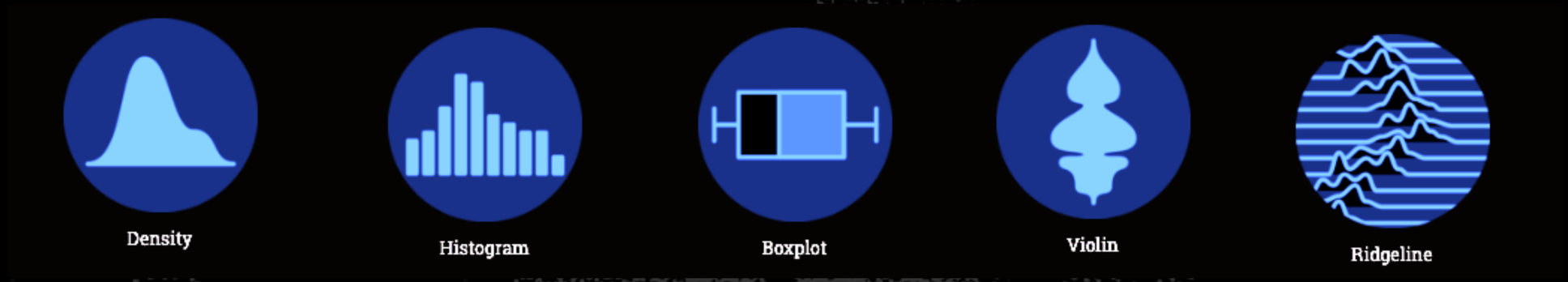
***Continuous data:***

***Not restricted to defined separate values***

***Can take any value over a continuous range***

***EX: normalized expression of a sequenced gene***

# DATA EXPLORATION




**RStudio, comes pre-loaded with example data-sets.**

# DATA EXPLORATION

## Diamonds Data Set

```
> str(diamonds)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       53940 obs. of  10 variables:
 $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
 $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
 $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
 $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
 $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
 $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y      : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z      : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

# DATA EXPLORATION

## Diamonds Data Set

```
> head(diamonds)
# A tibble: 6 x 10
  carat cut       color clarity depth table price     x     y     z
  <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
1 0.23  Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
2 0.21  Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
3 0.23  Good      E     VS1      56.9    65   327  4.05  4.07  2.31
4 0.290 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
5 0.31  Good      J     SI2      63.3    58   335  4.34  4.35  2.75
6 0.24  Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

# DATA EXPLORATION

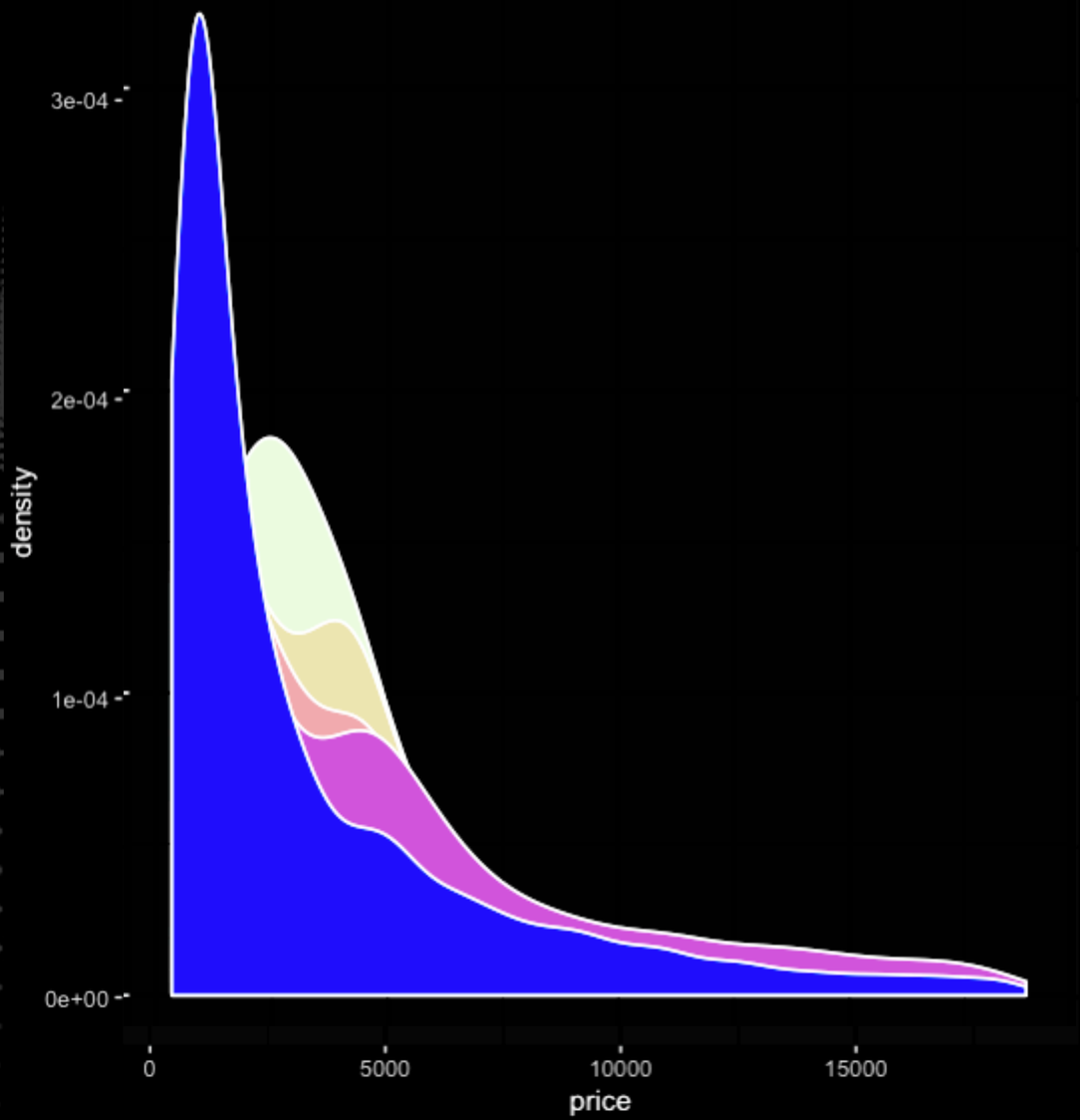
Density Plot

Allows to study the distribution of a *NUMERIC* continuous variable.

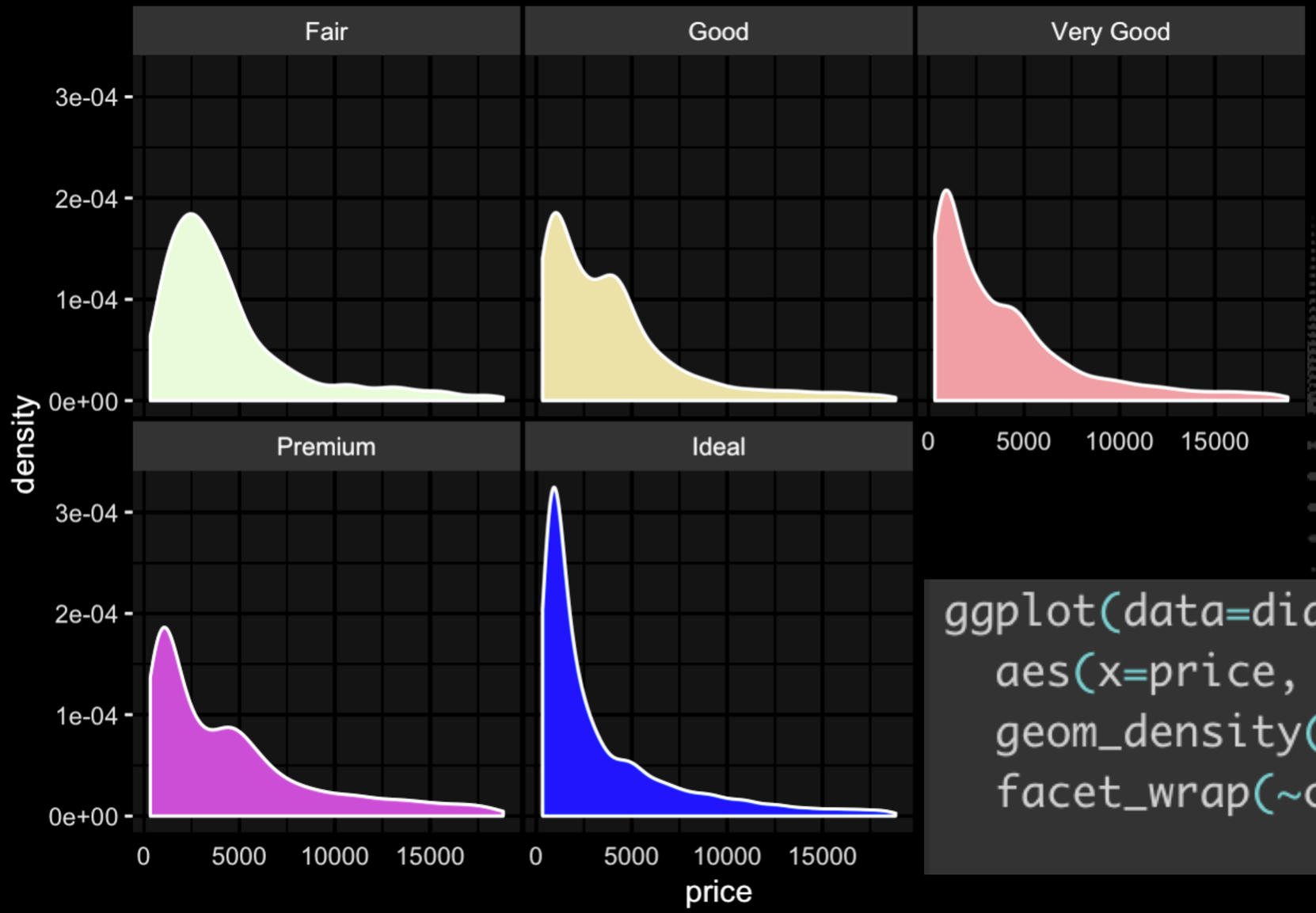
# DATA EXPLORATION

## Density Plot

```
library(ggplot2)
ggplot(data=diamonds,
  aes(x=price, group=cut, fill=cut)) +
  geom_density(adjust = 1.5, alpha = 0.4)
```



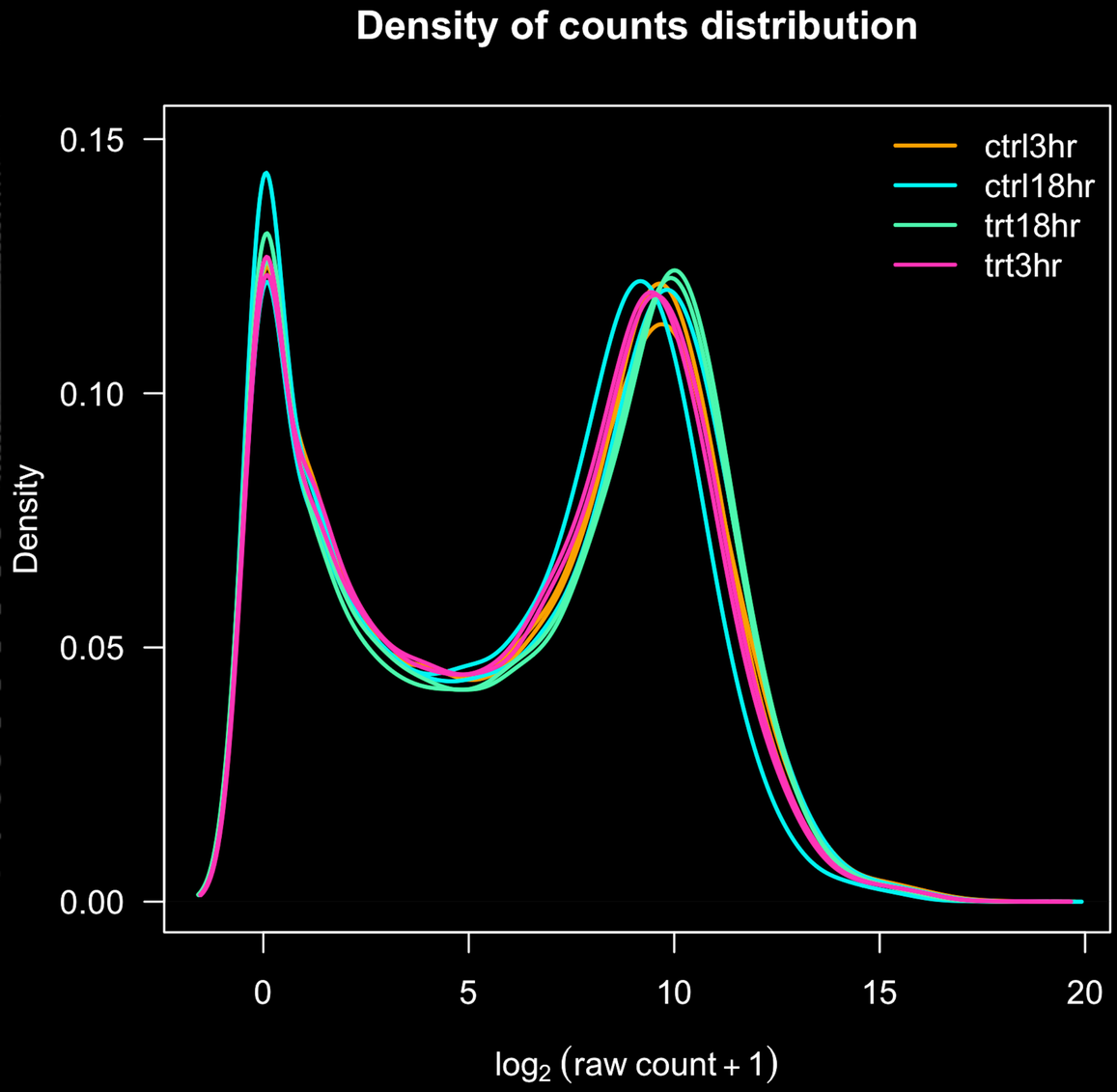
# DATA EXPLORATION




```
ggplot(data=diamonds,
  aes(x=price, group=cut, fill=cut)) +
  geom_density(adjust=1.5) +
  facet_wrap(~cut)
```

# IN CONTEXT OF RNA-seq

**Density Plot**



# DATA EXPLORATION

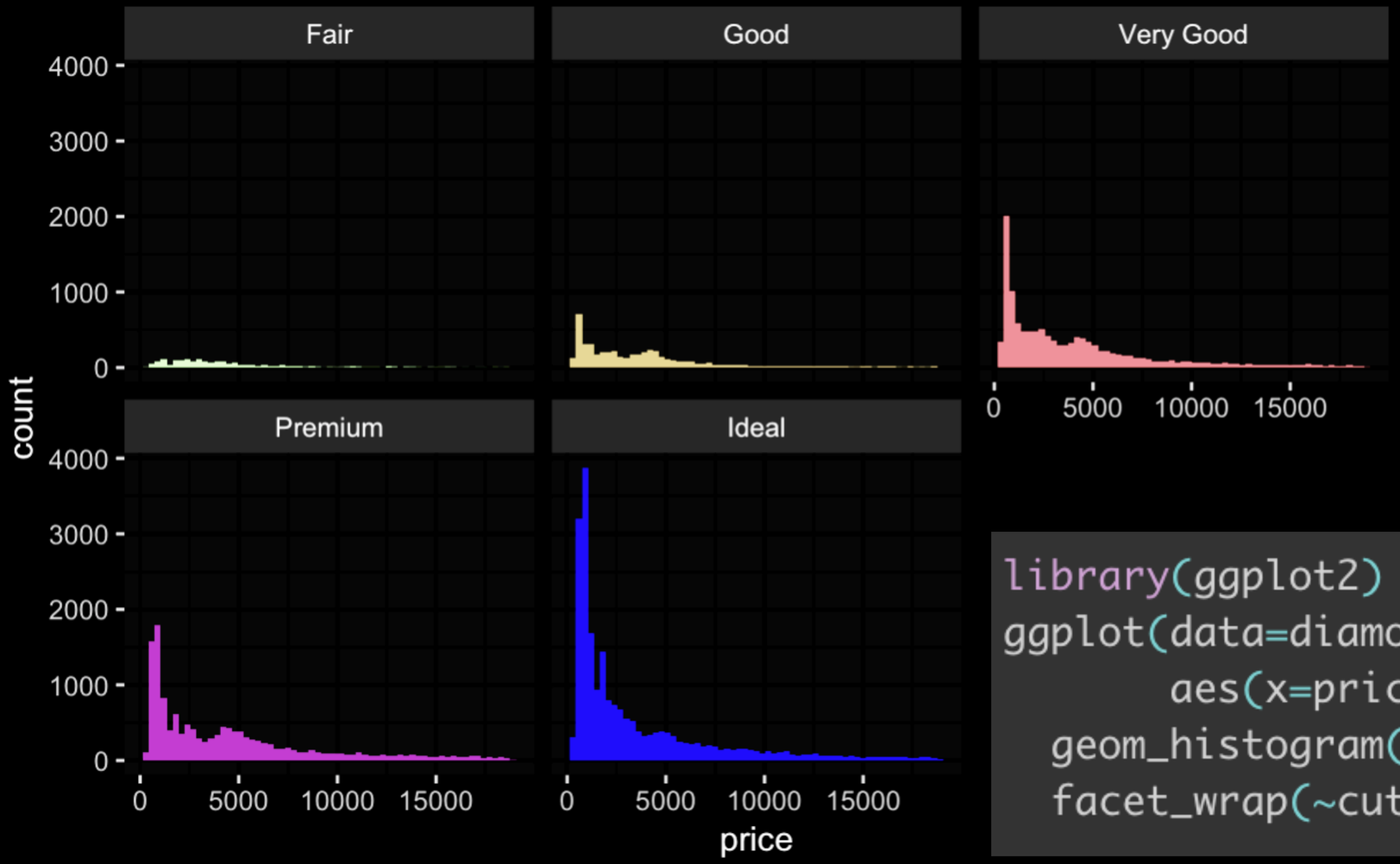
Histogram

Like Density Plot, allows us to study the distribution of a *NUMERIC* continuous variable.

The variable is cut into several bins, and the number of observation per bin is represented by the height of the bar.

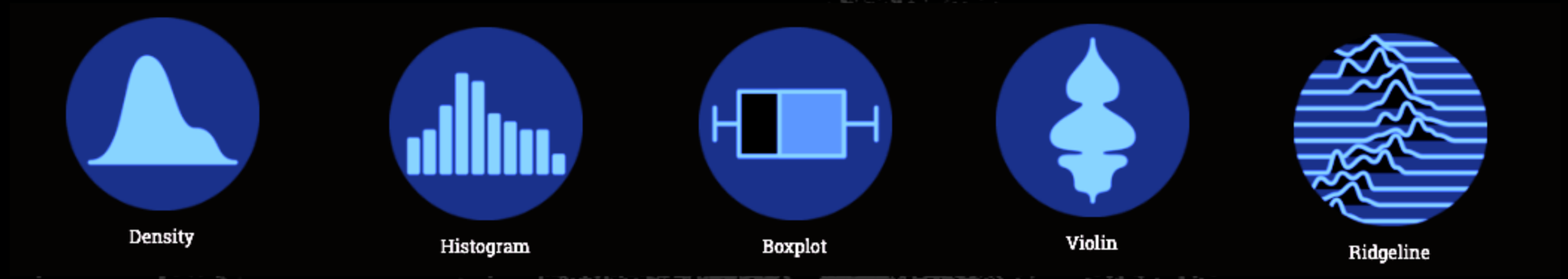
# DATA EXPLORATION

## Histogram

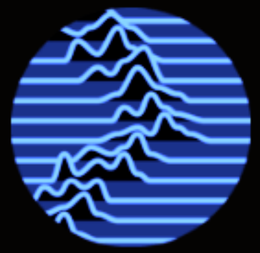



```
library(ggplot2)
ggplot(data=diamonds,
       aes(x=price, group=cut, fill=cut)) +
  geom_histogram(bins = 20,binwidth = 300) +
  facet_wrap(~cut)
```

# DATA EXPLORATION

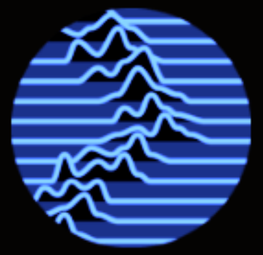


# DATA EXPLORATION



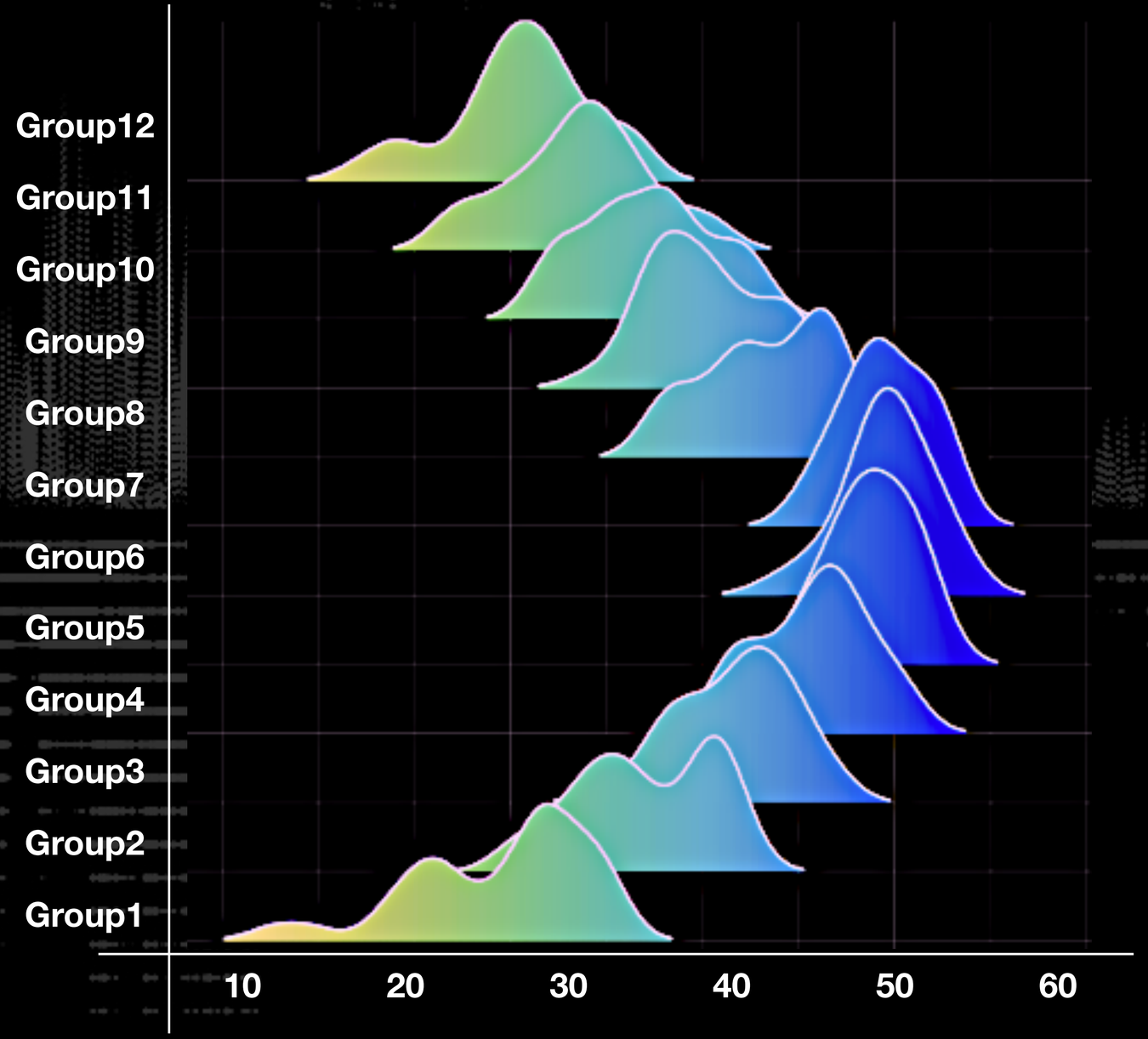
Ridgeline (Joyplot)

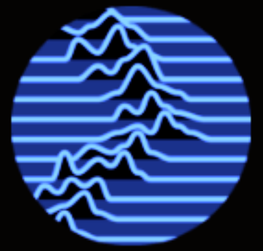
**Allows to study the distribution of a *NUMERIC* variable for several groups.**



# Ridgeline (Joyplot)

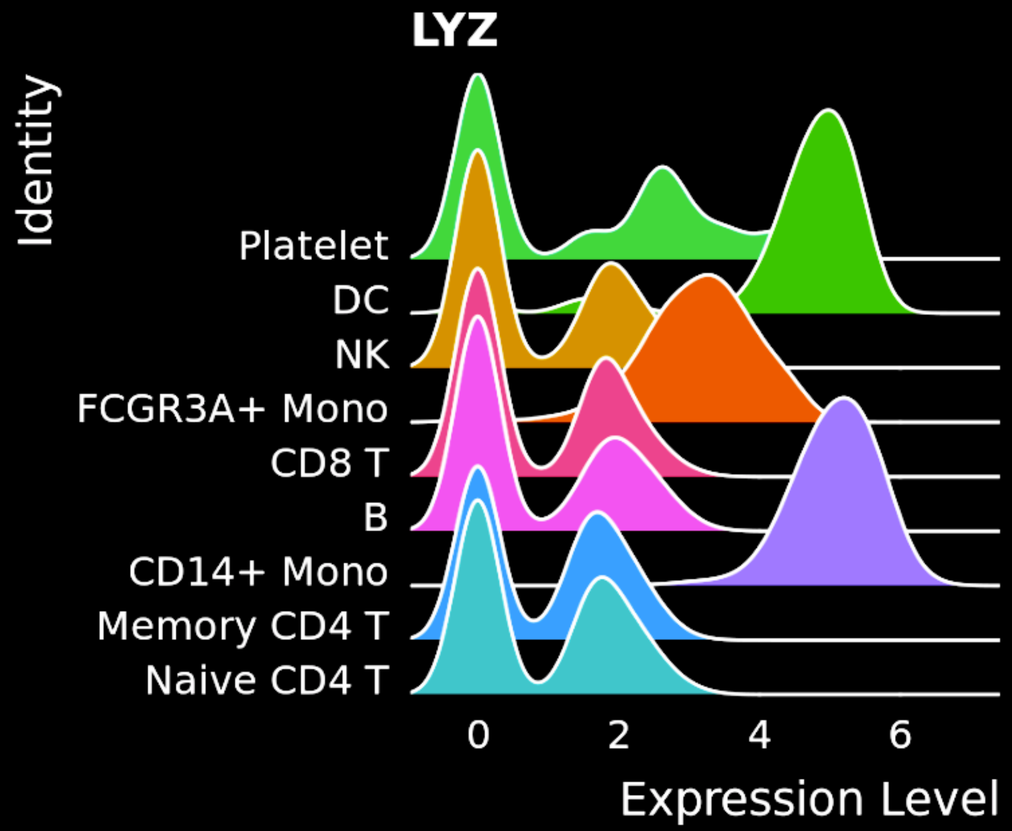
geom_density_ridges( )

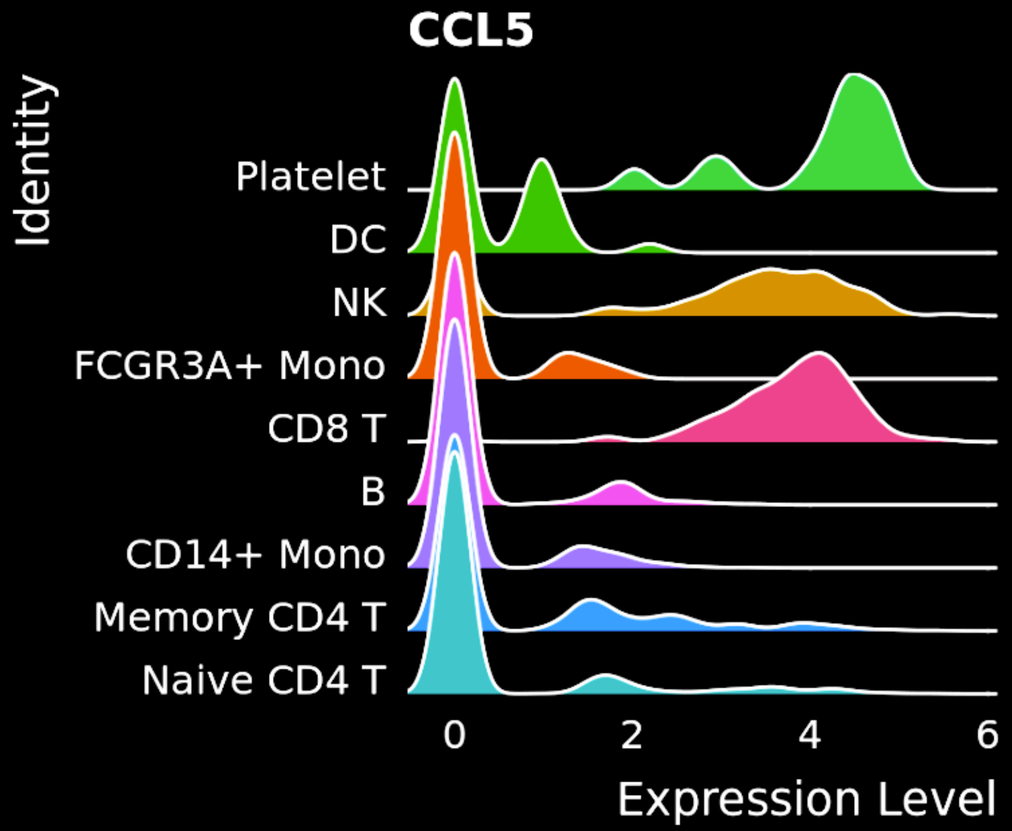


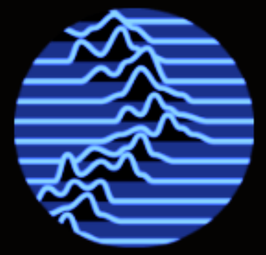


# Ridgeline (Joyplot)

## In Context of scRNA-seq:

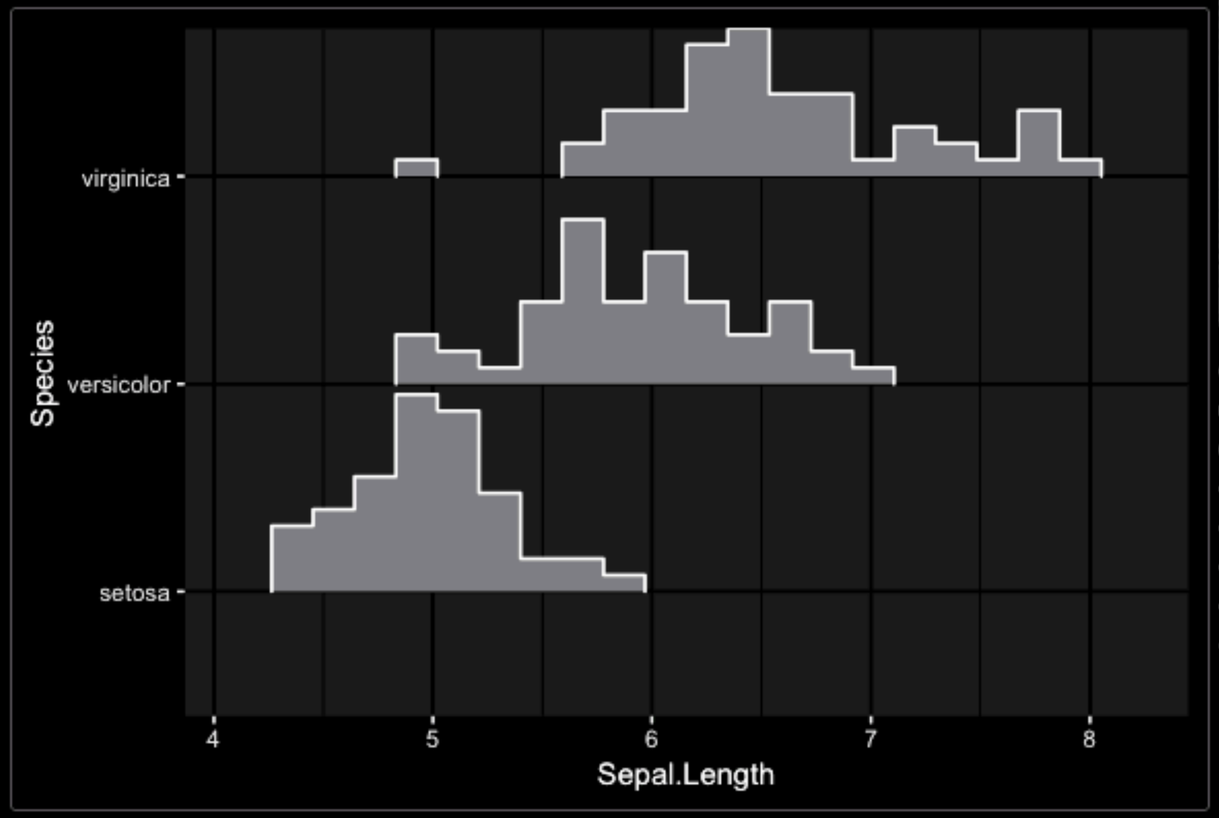


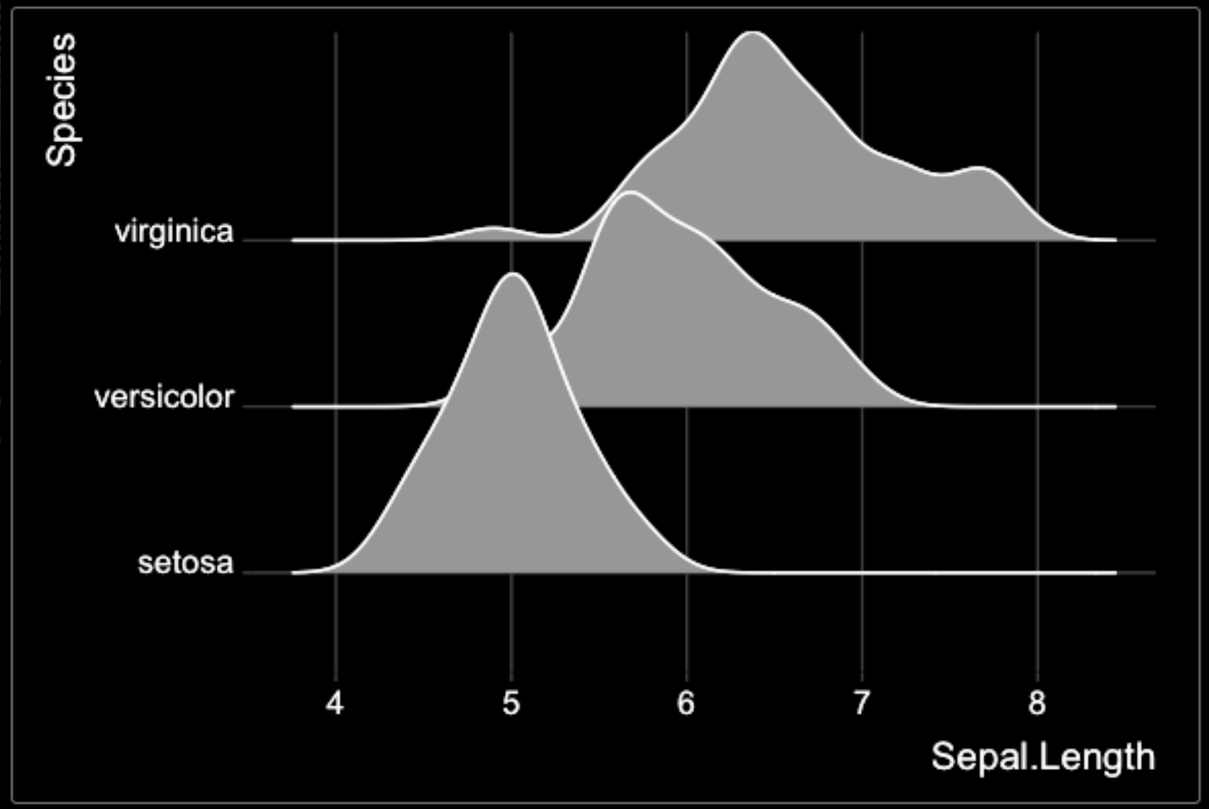





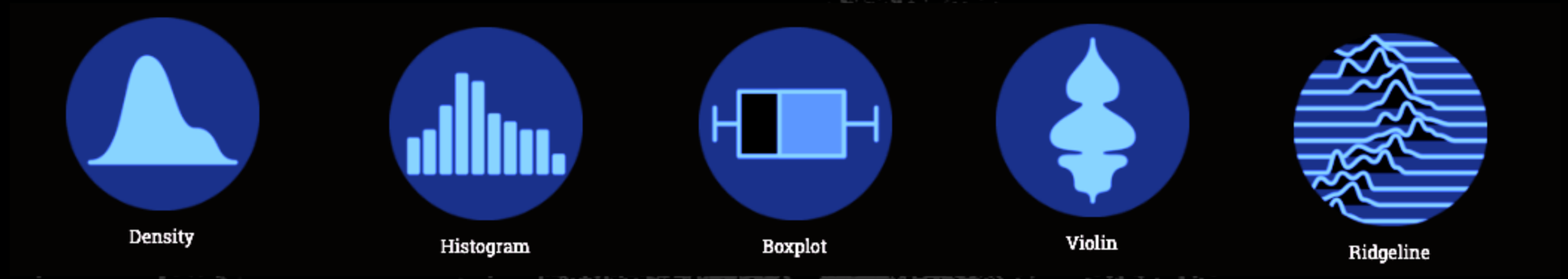
# Ridgeline (Joyplot)

## Variations:



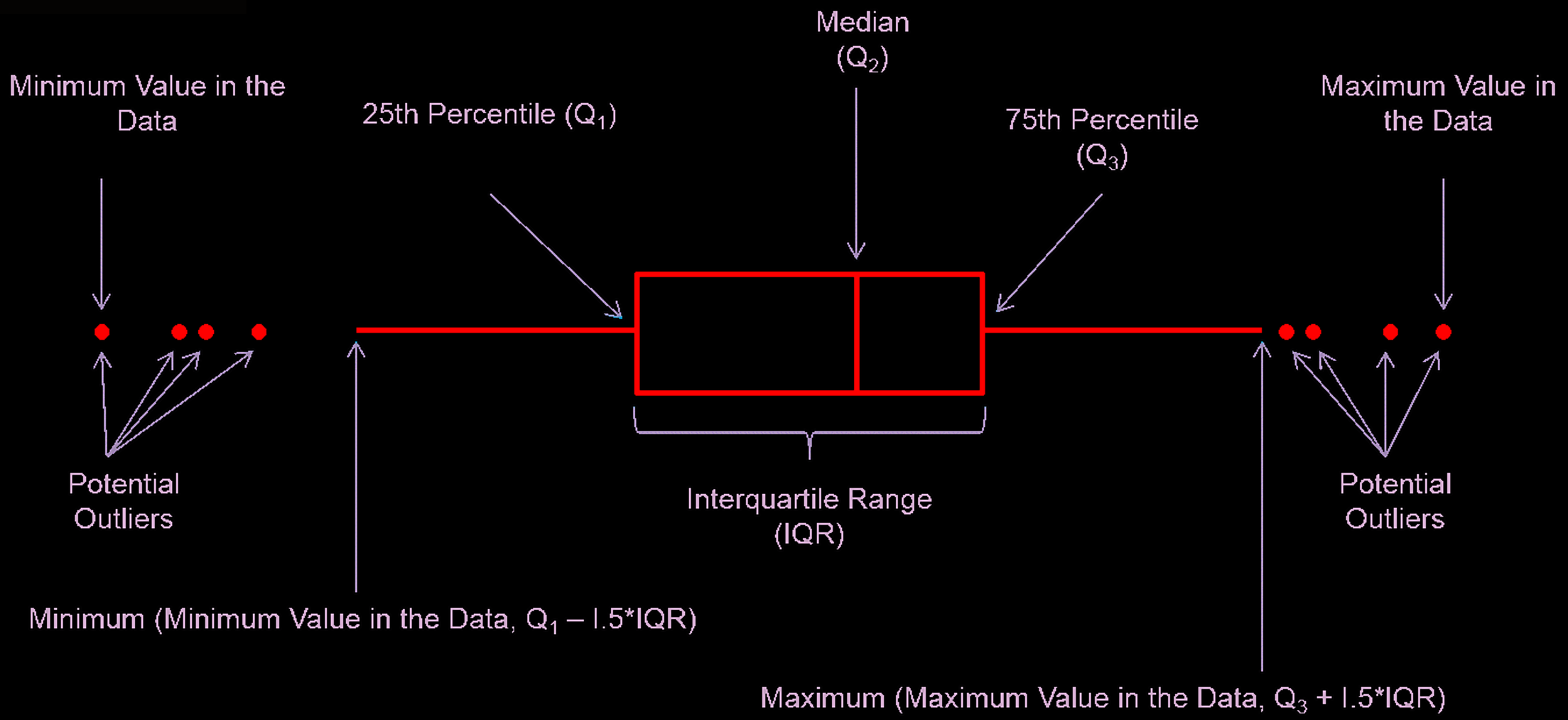



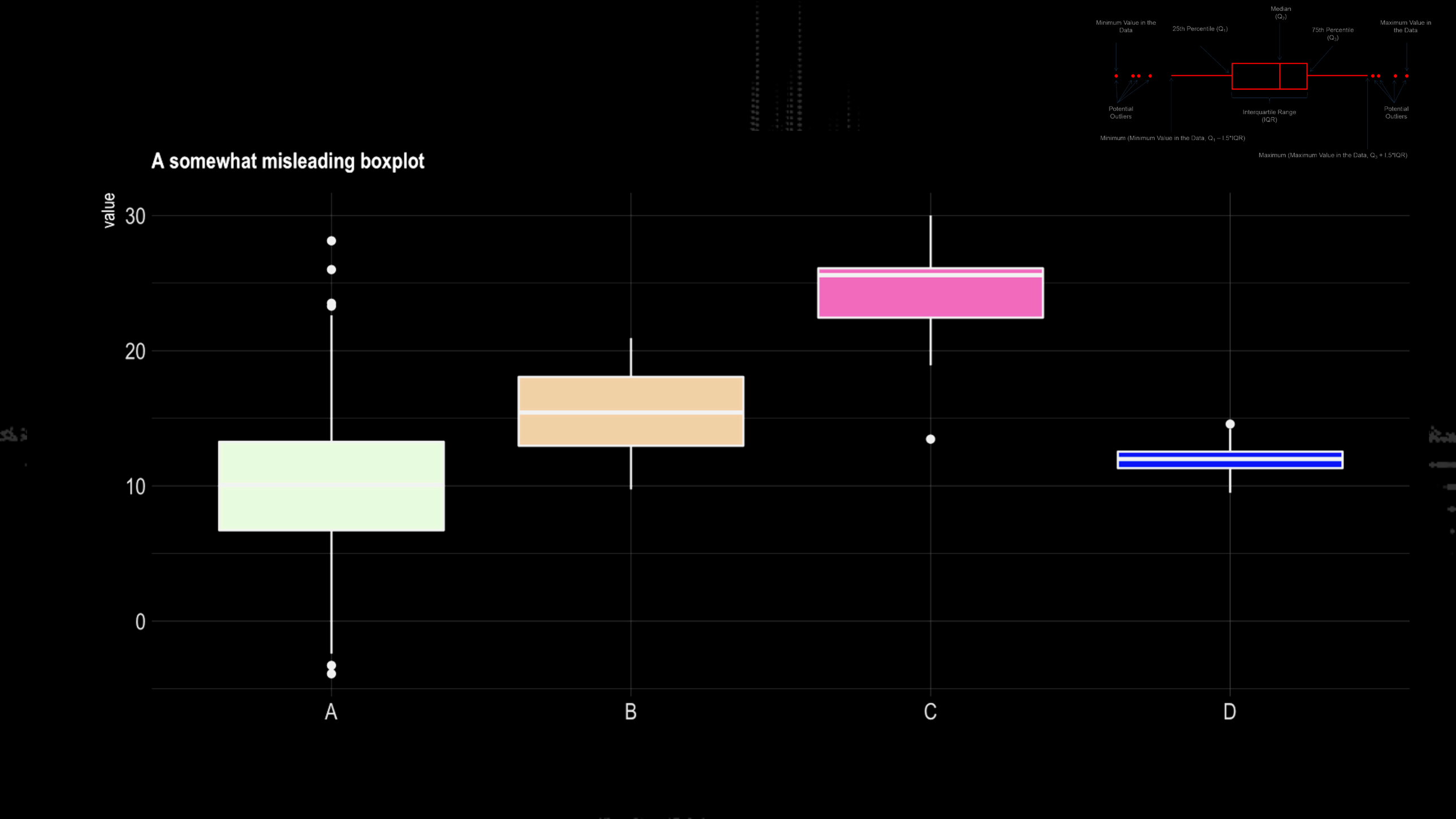
# DATA EXPLORATION

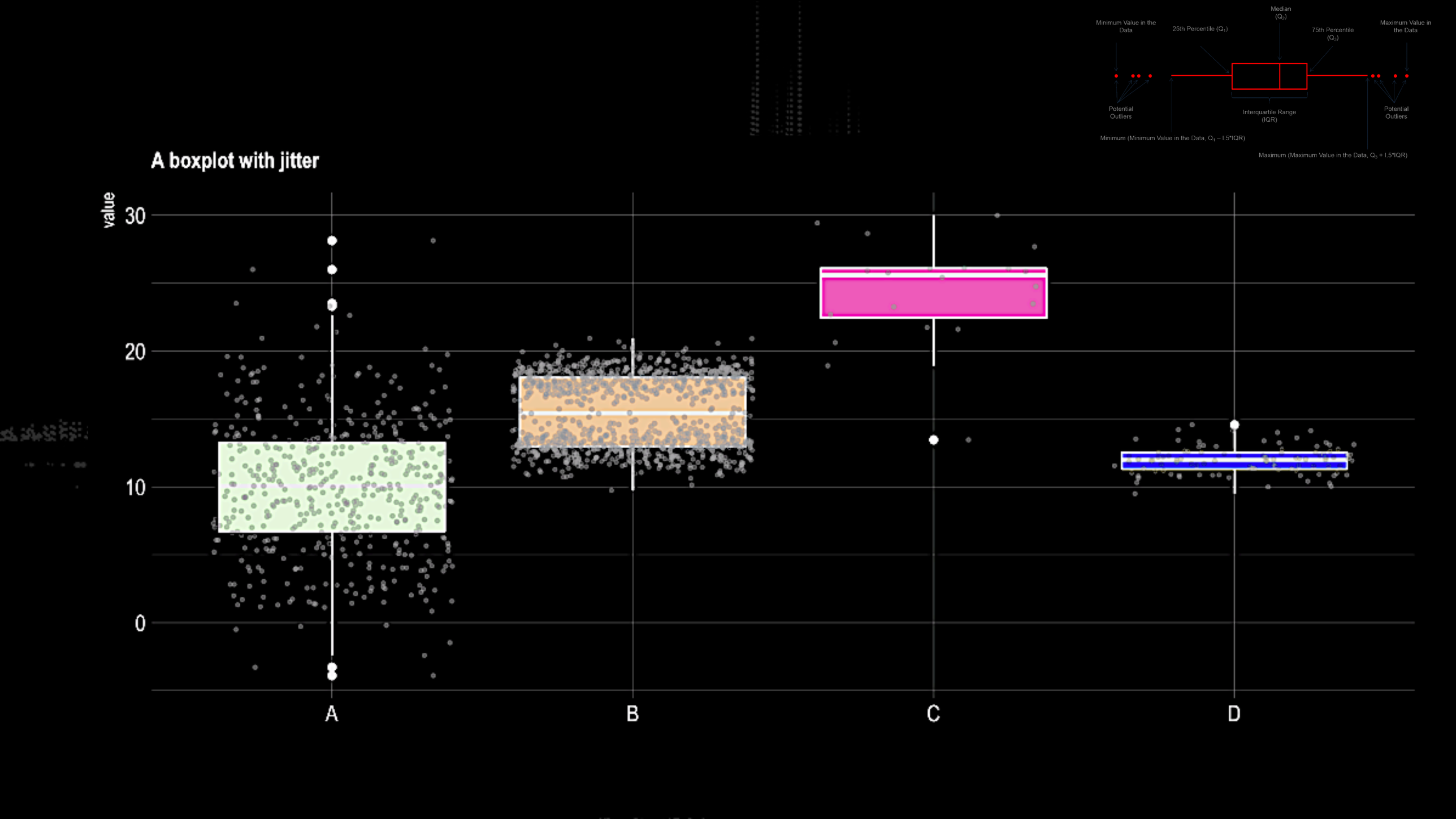


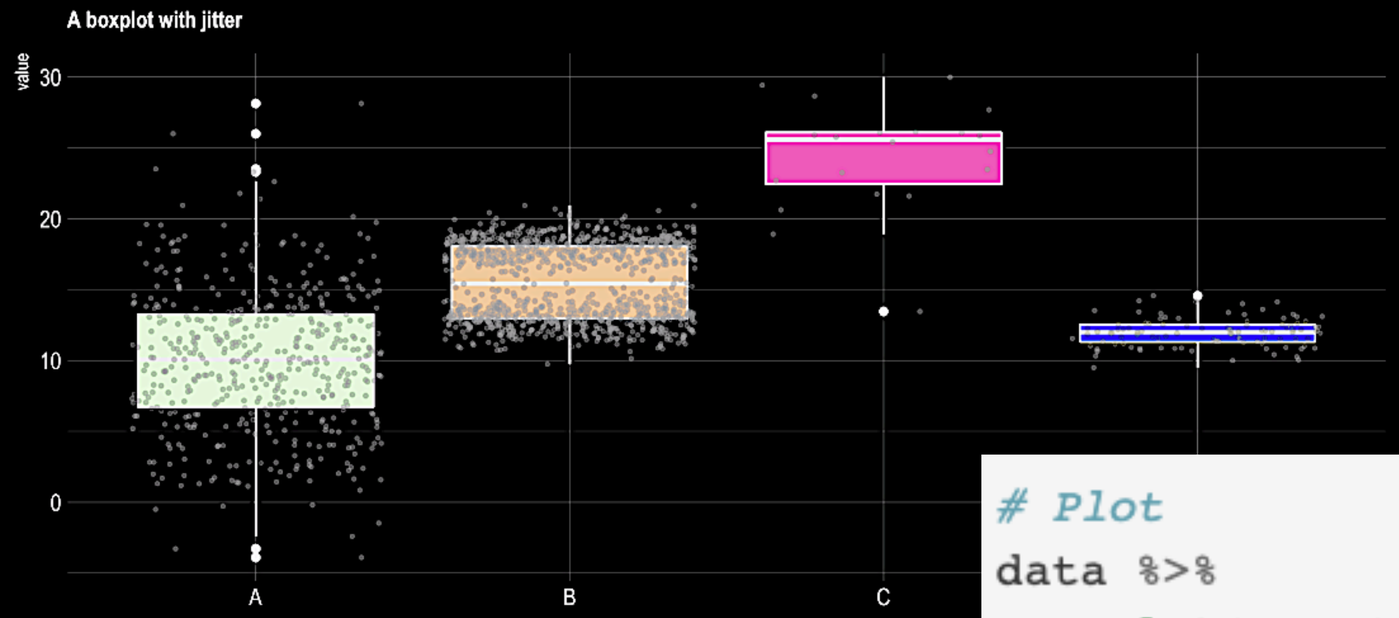

Boxplot

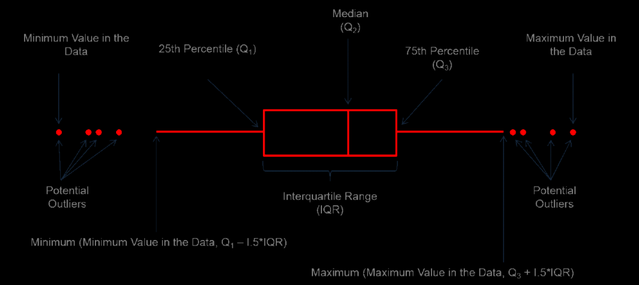
# Boxplots


Median
($Q_2$)
Minimum Value in the Data
25th Percentile ($Q_1$)
75th Percentile
($Q_3$)
Maximum Value in the Data
Potential Outliers
Interquartile Range
(IQR)
Potential Outliers
Minimum (Minimum Value in the Data, $Q_1$ – 1.5*IQR)
Maximum (Maximum Value in the Data, $Q_3$ + 1.5*IQR)


A somewhat misleading boxplot
value
30
20
10
0
A
B
C
D
Median ($Q_2$)
Minimum Value in the Data
25th Percentile ($Q_1$)
75th Percentile ($Q_3$)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Minimum (Minimum Value in the Data, $Q_1 - 1.5*IQR$)
Maximum (Maximum Value in the Data, $Q_3 + 1.5*IQR$)

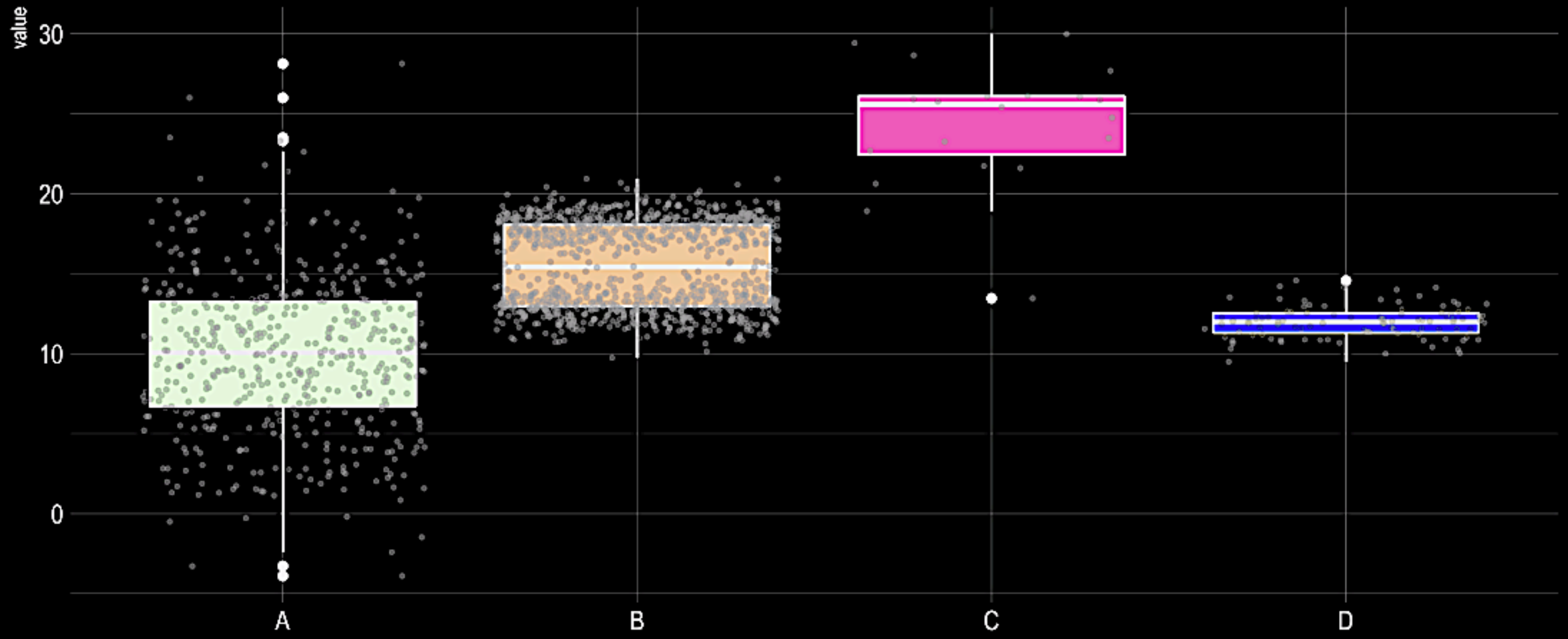

A boxplot with jitter
value
30
20
10
0
A
B
C
D
Minimum Value in the Data
25th Percentile ($Q_1$)
Median ($Q_2$)
75th Percentile ($Q_3$)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Minimum (Minimum Value in the Data, $Q_1 - 1.5*IQR$)
Maximum (Maximum Value in the Data, $Q_3 + 1.5*IQR$)

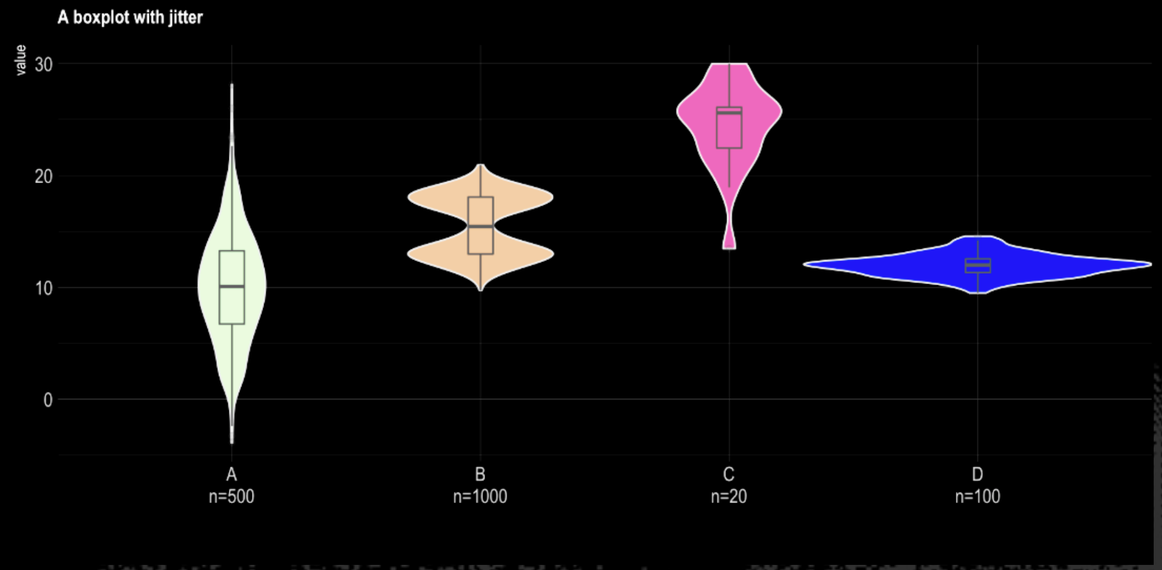



```r
# Plot
data %>%
  ggplot( aes(x=name, y=value, fill=name)) +
    geom_boxplot() +
    scale_fill_viridis(discrete = TRUE) +
    geom_jitter(color="grey", size=0.7, alpha=0.5) +
    theme_ipsum() +
    theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    ggtitle("A boxplot with jitter") +
    xlab("")
```

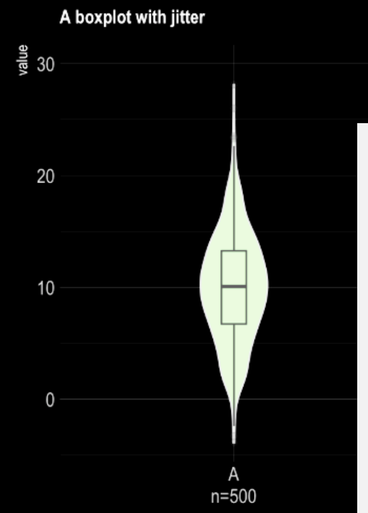
# Violin Plot






```r
# sample size
sample_size = data %>% group_by(name) %>% summarize(num=n())

# Plot
data %>%
  left_join(sample_size) %>%
  mutate(myaxis = paste0(name, "\n", "n=", num)) %>%
  ggplot( aes(x=myaxis, y=value, fill=name)) +
    geom_violin(width=1.4) +
    geom_boxplot(width=0.1, color="grey", alpha=0.2) +
    scale_fill_viridis(discrete = TRUE) +
    theme_ipsum() +
    theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    ggtitle("A boxplot with jitter") +
    xlab("")
```

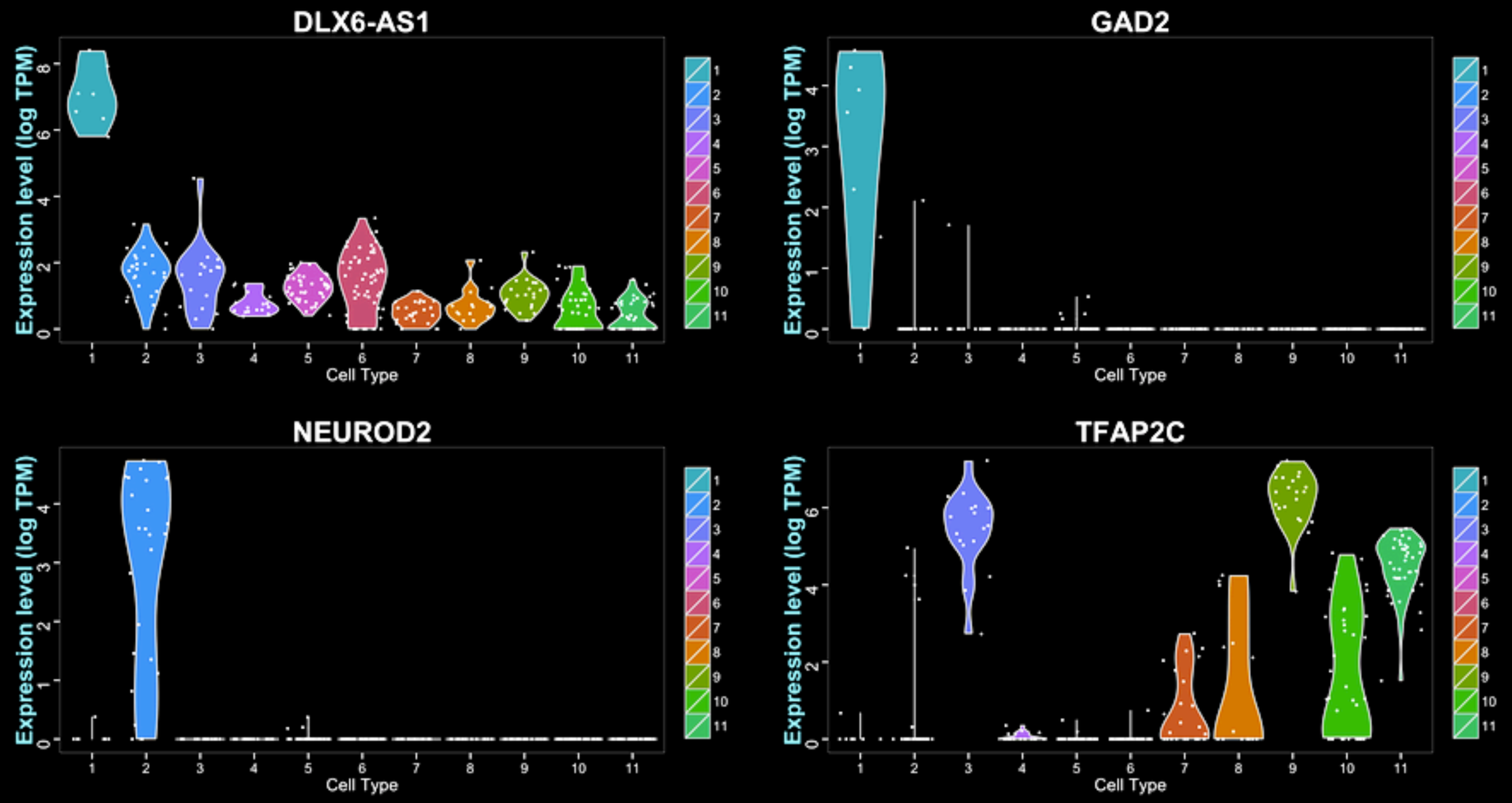



```r
# sample size
sample_size = data %>% group_by(name) %>% summarize(num=n())

# Plot
data %>%
  left_join(sample_size) %>%
  mutate(myaxis = paste0(name, "\n", "n=", num)) %>%
  ggplot( aes(x=myaxis, y=value, fill=name)) +
    geom_violin(width=1.4) +
    geom_boxplot(width=0.1, color="grey", alpha=0.2) +
    scale_fill_viridis(discrete = TRUE) +
    theme_ipsum() +
    theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    ggtitle("A boxplot with jitter") +
    xlab("")
```

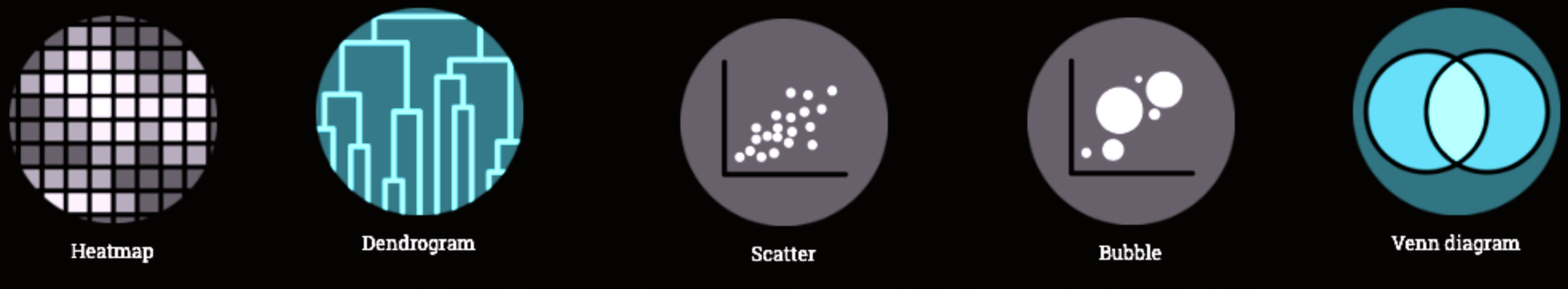
# IN CONTEXT OF scRNA-seq

## Violin Plot



# WHAT IS MY GOAL?

- **DATA SUMMARY?**



# DATA EXPLORATION

Heatmap

A ***heatmap*** is a graphical representation of data, where the individual values contained in a matrix are represented as colors.

Heatmap

# Mtcars Data Set

```
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Heatmap

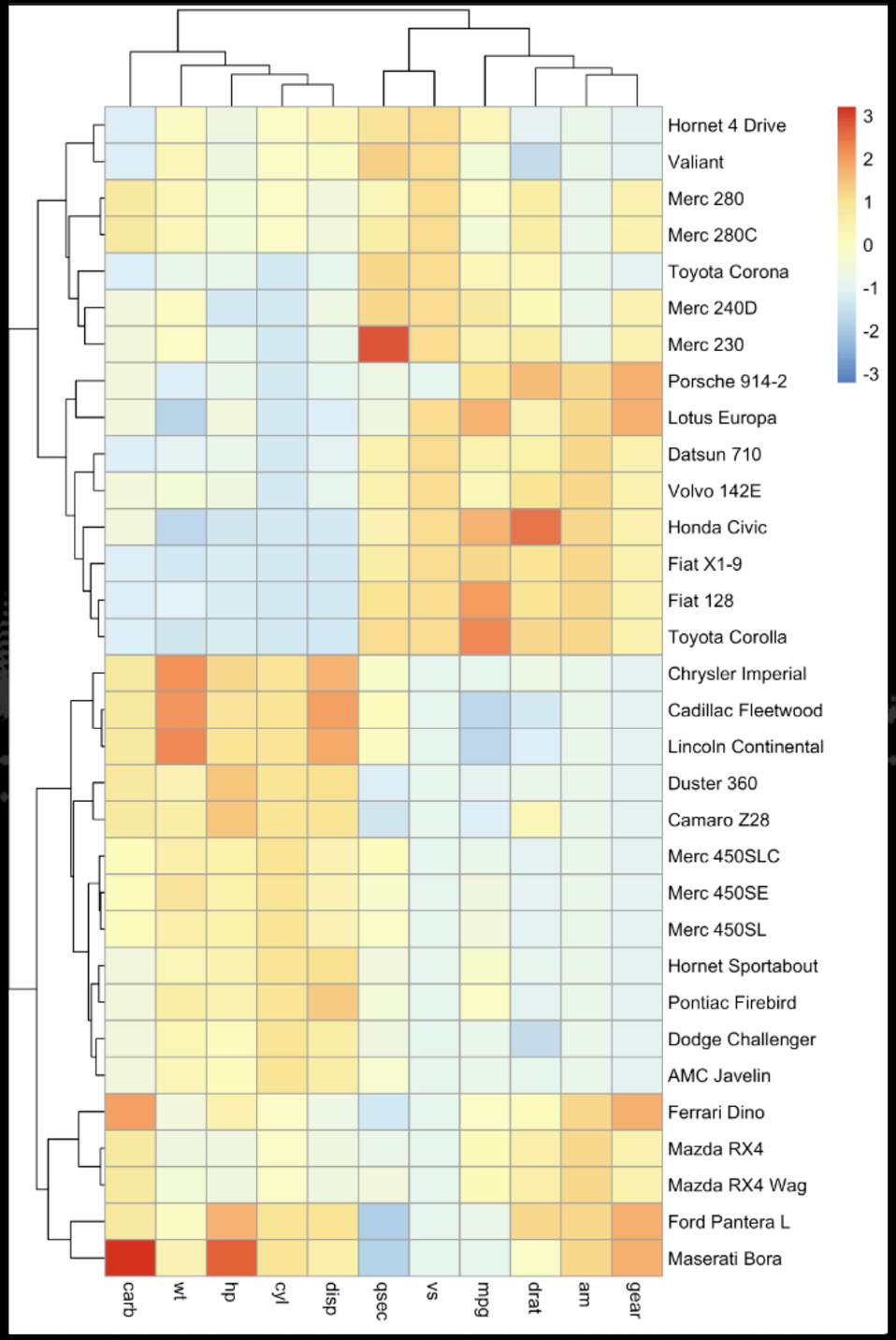
# Mtcars Data Set

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

# Heatmap

## Mtcars Data Set

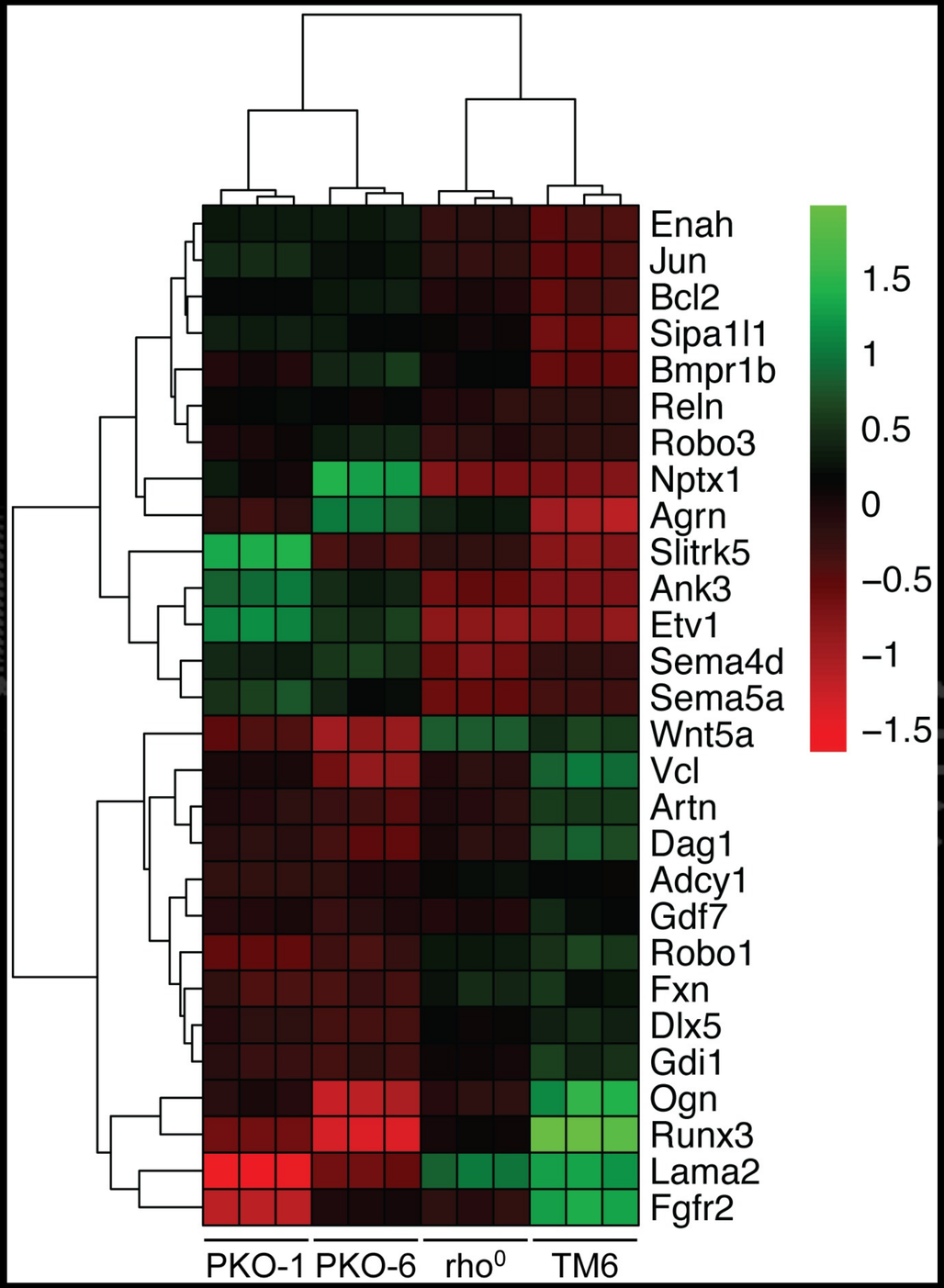
```
library(pheatmap)
pheatmap(mat = mtcars, scale = "column")
```



# Heatmap

## In Context of RNA-seq:

| | wt1 | wt2 | wt3 | ko1 | ko2 | ko3 |
|---|---|---|---|---|---|---|
| gene1 | 135 | 148 | 146 | 269 | 268 | 227 |
| gene2 | 803 | 797 | 841 | 412 | 408 | 388 |
| gene3 | 40 | 25 | 38 | 413 | 393 | 417 |
| gene4 | 381 | 383 | 415 | 809 | 840 | 859 |
| gene5 | 775 | 766 | 773 | 302 | 310 | 324 |
| gene6 | 305 | 313 | 256 | 831 | 817 | 832 |
| gene7 | 816 | 819 | 800 | 485 | 481 | 429 |
| gene8 | 40 | 22 | 40 | 421 | 476 | 479 |
| gene9 | 963 | 935 | 938 | 43 | 26 | 41 |
| gene10 | 697 | 749 | 715 | 233 | 259 | 284 |
| gene11 | 36 | 50 | 40 | 168 | 178 | 168 |
| gene12 | 60 | 66 | 54 | 288 | 289 | 293 |
| gene13 | 537 | 517 | 523 | 142 | 134 | 145 |



# DATA EXPLORATION

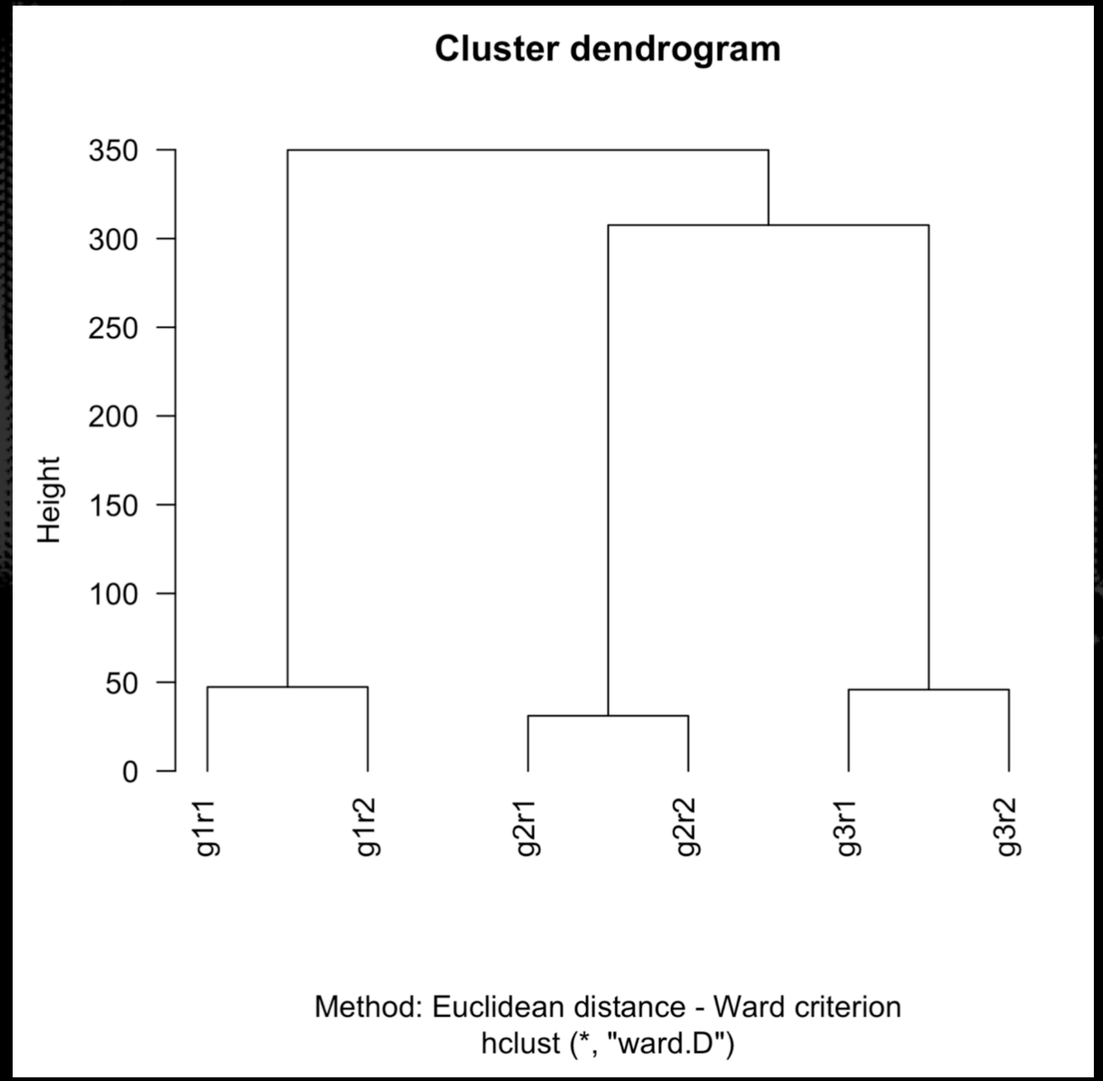
Dendrogram

A network structure consisting of nodes and edges

# DATA EXPLORATION

**Dendrogram**

```
library(DESeq2)
hc2 <- hclust(dist(t(assay(vsd))),
              method="ward.D")
plot(hc2, hang=-1, ylab="Height",
xlab="Method: Euclidean distance
- Ward criterion",
main="Cluster Dendrogram")
```



# DATA EXPLORATION

Scatter Plot

Displays a relationship between two ***NUMERICAL*** variables

Scatter Plot

# Iris Data Set

```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1
 1 1 1 1 1 ...
```

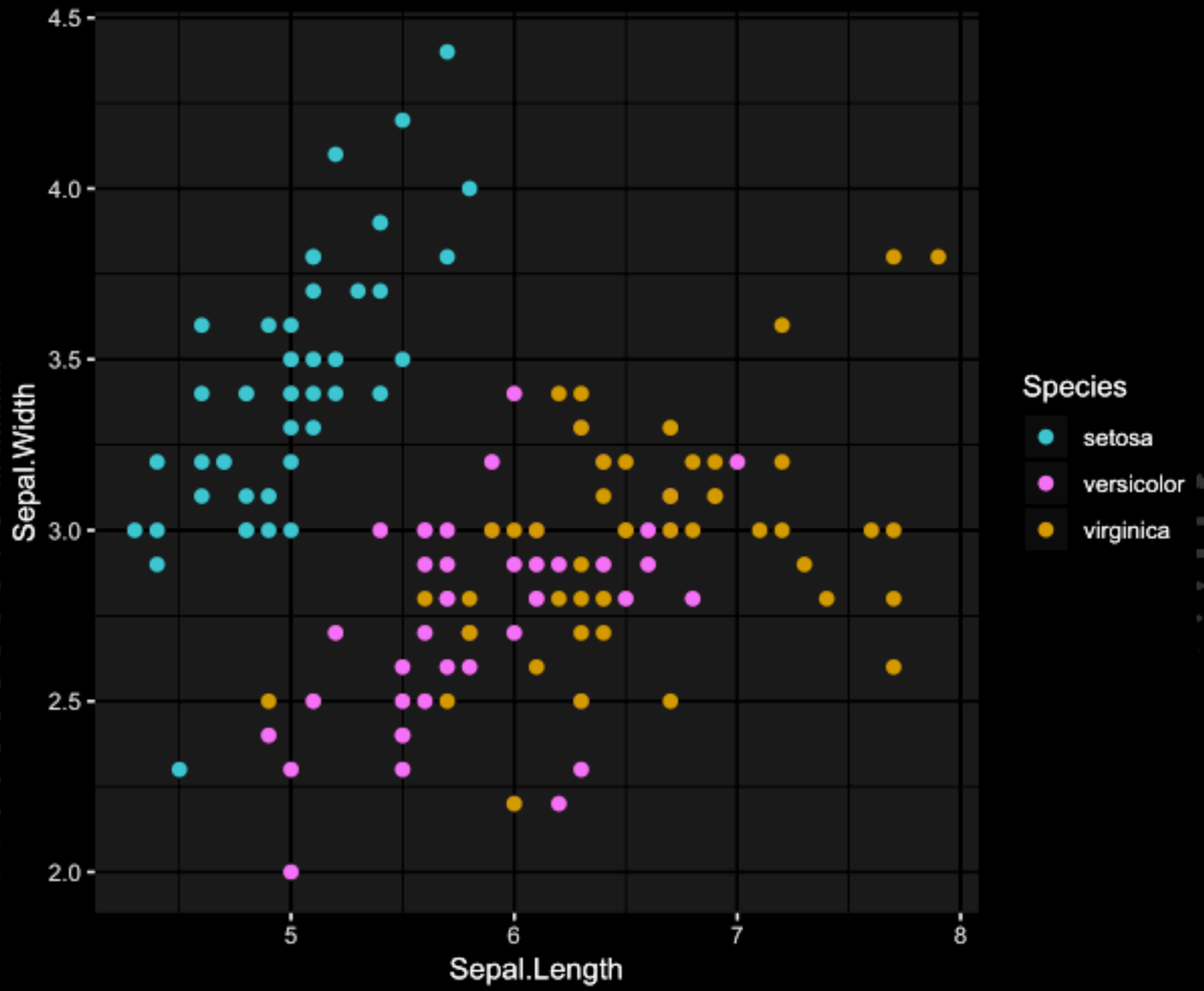
Scatter Plot

# Iris Data Set

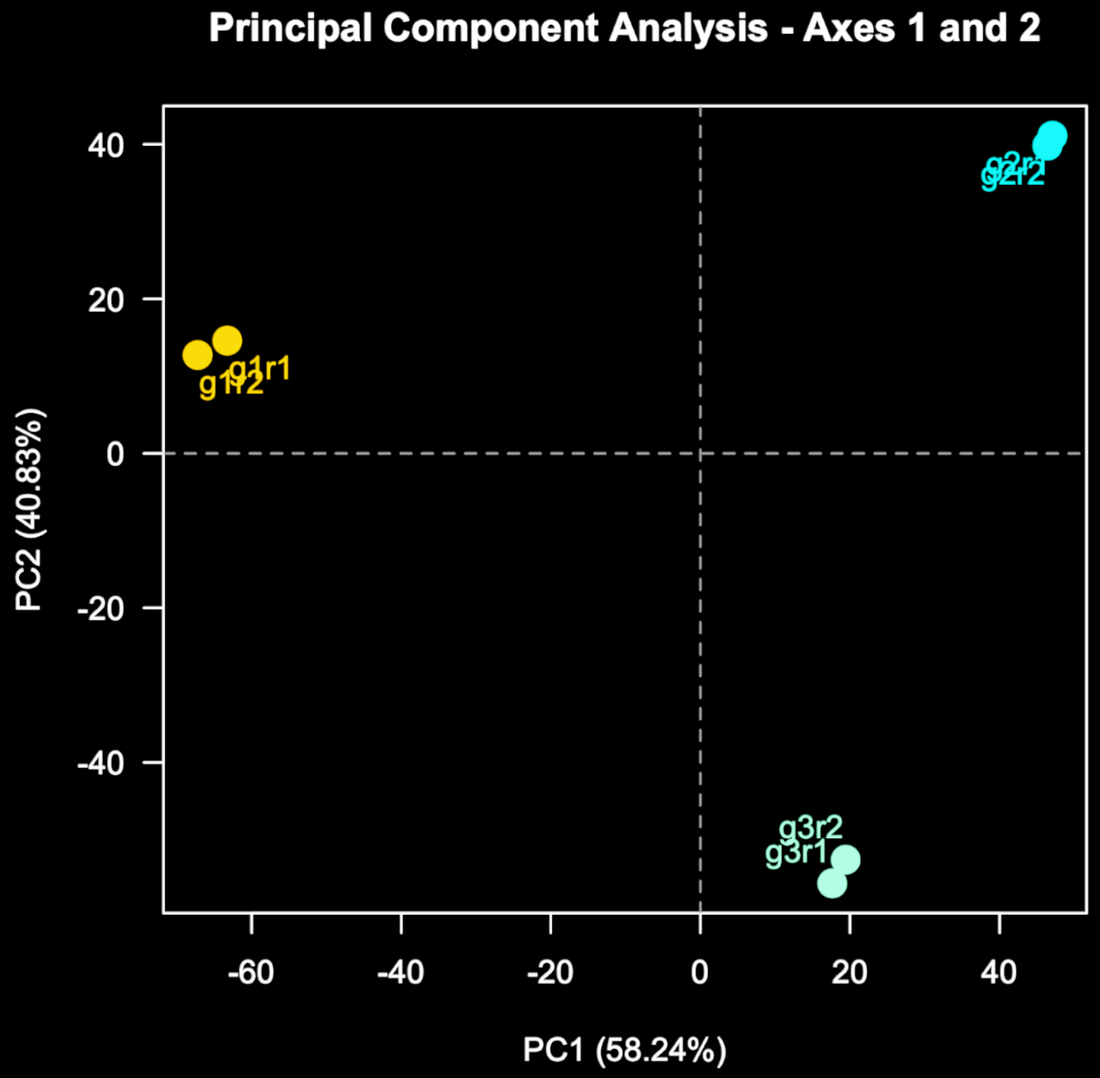
```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

# Scatter Plot

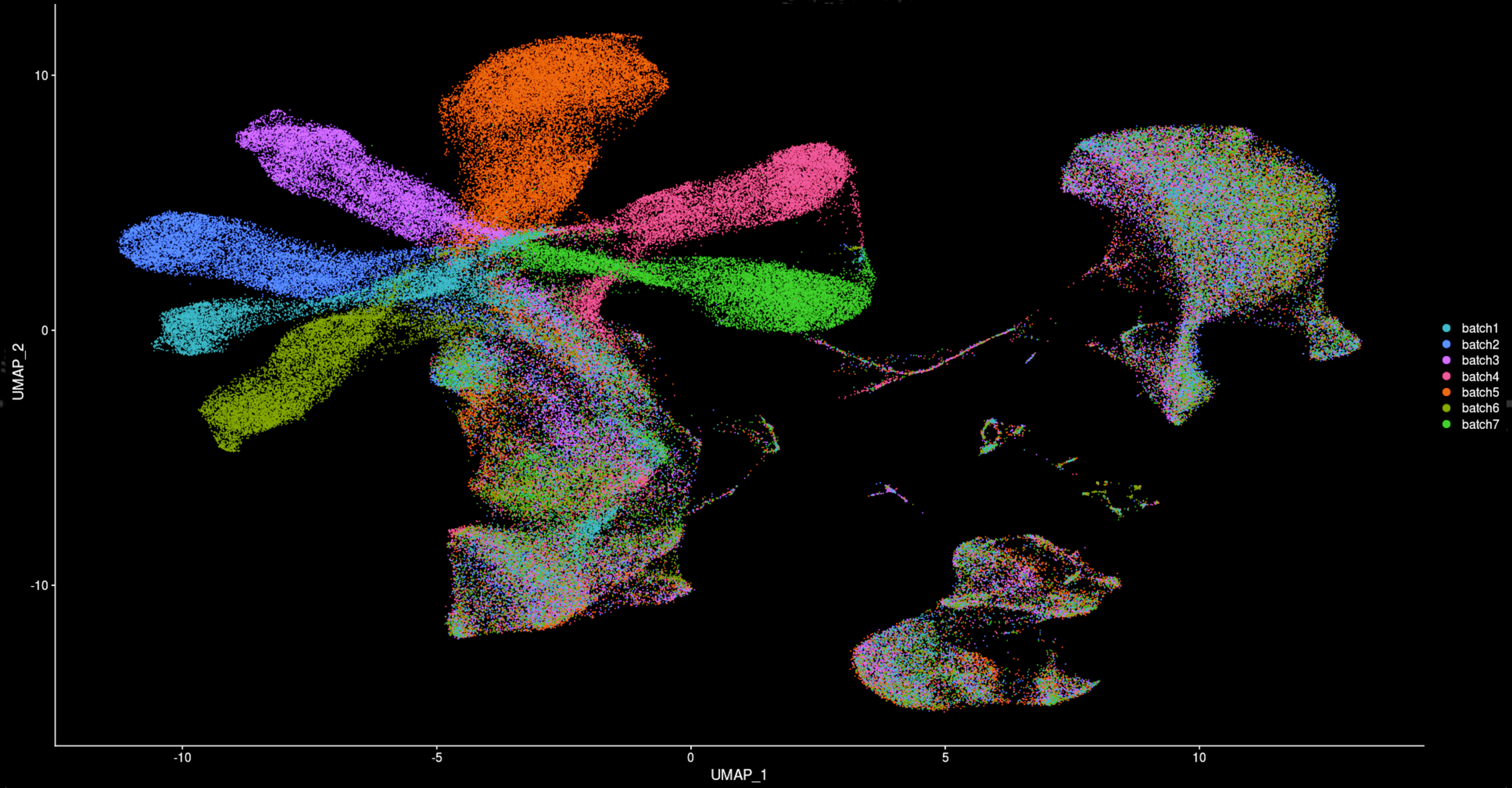
```
library(ggplot2)
ggplot(iris, aes(x=Sepal.Length,
  y=Sepal.Width, color=Species)) +
  geom_point(size=2)
```



# IN CONTEXT OF RNA-seq



# IN CONTEXT OF scRNA-seq

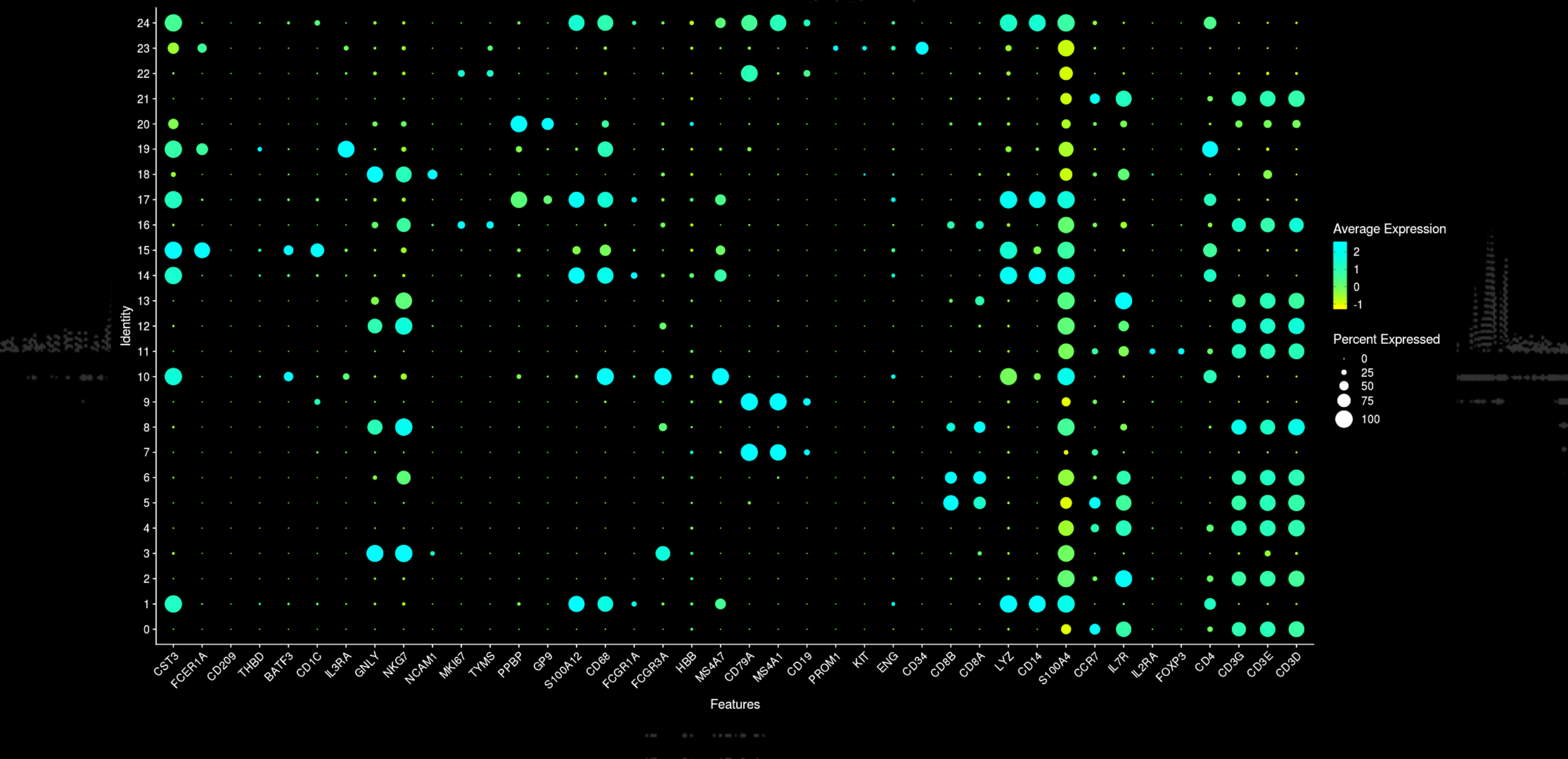


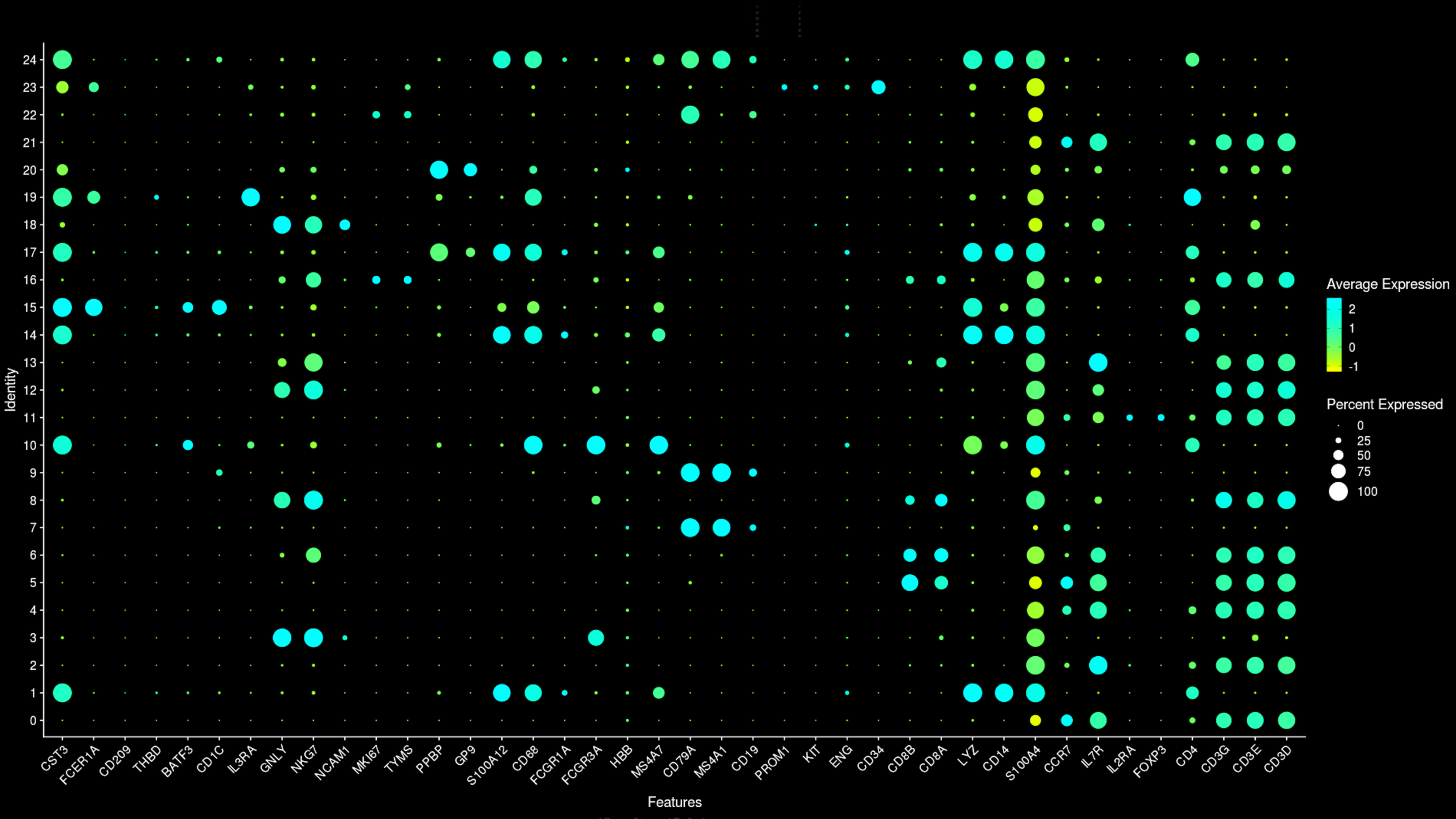
# DATA EXPLORATION

Bubble Plot

A ***bubble plot*** is a ***scatterplot*** where a third dimension is added: the value of an additional numeric variable is represented through the size of the dots.

# IN CONTEXT OF scRNA-seq




Identity
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
CST3
FCER1A
CD209
THBD
BATF3
CD1C
IL3RA
GNLY
NKG7
NCAM1
MKI67
TYMS
PPBP
GP9
S100A12
CD68
FCGR1A
FCGR3A
HBB
MS4A7
CD79A
MS4A1
CD19
PROM1
KIT
ENG
CD34
CD8B
CD8A
LYZ
CD14
S100A4
CCR7
IL7R
IL2RA
FOXP3
CD4
CD3G
CD3E
CD3D
Features
Average Expression
2
1
0
-1
Percent Expressed
0
25
50
75
100

# DATA EXPLORATION

VennDiagram Plot

A *Venn diagram* shows all possible logical relationships between a finite collection of different sets.

# VennDiagram Plot

# UPSET PLOT