McGill initiative in Computational Medicine

McGill

Data Visualization in R: Essentials and Optimization

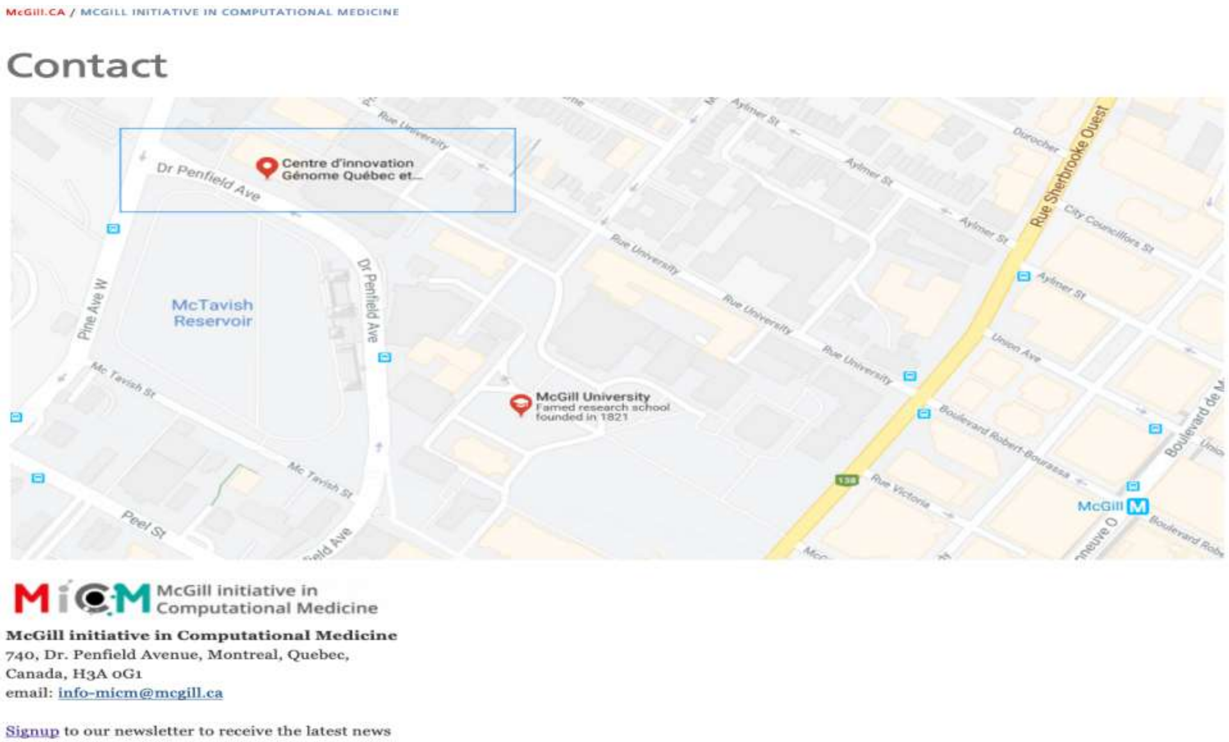
Instructor: Octavia Maria Dancu-Lixandru

Pre-Intro Prep

- Before we get started with the course intro and material, let's make sure:
  - your computer is set up with R & RStudio correctly installed and open
  - You have downloaded the datasets
  - You have loaded the data into your R session
  - Open this google document: https://bit.ly/2AOI04A
  - This is where we'll be sharing the plots made in class

MiCM McGill initiative in Computational Medicine

<u>Mission</u> : aims to deliver inter-disciplinary research programs and empower the use of data in health research and health care delivery



McGill.CA / MCGILL INITIATIVE IN COMPUTATIONAL MEDICINE

Contact

MiCM McGill initiative in Computational Medicine

McGill initiative in Computational Medicine
740, Dr. Penfield Avenue, Montreal, Quebec,
Canada, H3A 0G1
email: info-micm@mcgill.ca

Signup to our newsletter to receive the latest news

https://www.mcgill.ca/micm

About Me:

- Octavia M. Dancu-Lixandru
- Human Genetics, MSc 2, Majewski Lab
- Epigenetics of Head & Neck Cancer (methylation data, expression data, mutation data…)

What is your R background?

Why are you taking this course?

About You:

- What kind of data do you normally handle?
- How do you usually make figures and where are you typically using figures?

Intro to Data Visualization:

- 2 Main Purposes of Data Visualization:

  1. Presentation

  2. Exploration

Presentation

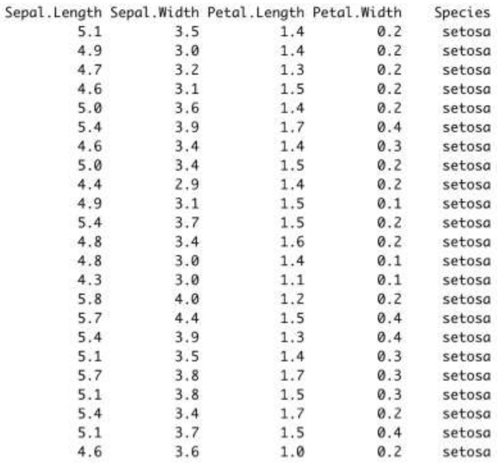
Presentation

- Communicate your ideas and features of your data
- Facilitate Understanding
- Transparency

Exploration

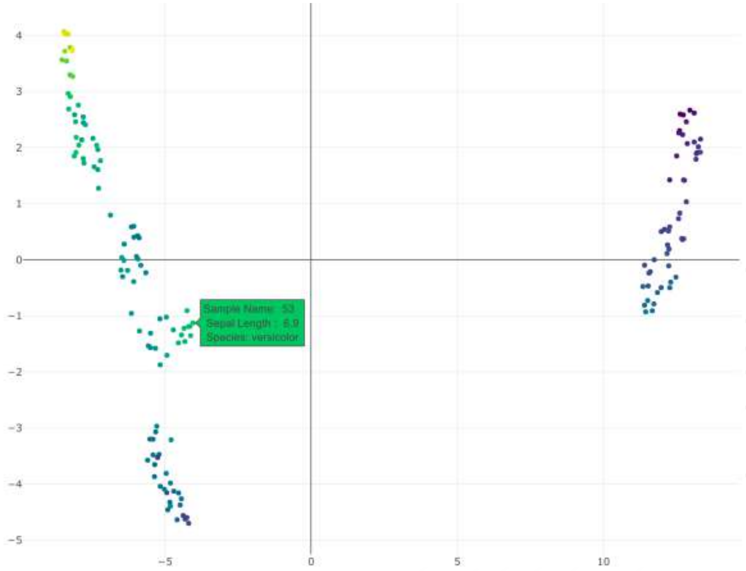
Exploration

- Convert data from tables (non-intuitive) ….



```
Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
         5.1         3.5          1.4         0.2     setosa
         4.9         3.0          1.4         0.2     setosa
         4.7         3.2          1.3         0.2     setosa
         4.6         3.1          1.5         0.2     setosa
         5.0         3.6          1.4         0.2     setosa
         5.4         3.9          1.7         0.4     setosa
         4.6         3.4          1.4         0.3     setosa
         5.0         3.4          1.5         0.2     setosa
         4.4         2.9          1.4         0.2     setosa
         4.9         3.1          1.5         0.1     setosa
         5.4         3.7          1.5         0.2     setosa
         4.8         3.4          1.6         0.2     setosa
         4.8         3.0          1.4         0.1     setosa
         4.3         3.0          1.1         0.1     setosa
         5.8         4.0          1.2         0.2     setosa
         5.7         4.4          1.5         0.4     setosa
         5.4         3.9          1.3         0.4     setosa
         5.1         3.5          1.4         0.3     setosa
         5.7         3.8          1.7         0.3     setosa
         5.1         3.8          1.5         0.3     setosa
         5.4         3.4          1.7         0.2     setosa
         5.1         3.7          1.5         0.4     setosa
         4.6         3.6          1.0         0.2     setosa
```

- To:
  - Graphical displays
  - Statistical models



Exploration

- Gain further information
- Generate hypotheses
- Confirm or discard theories
- Analyze your data

Goals of this course:

- By the end of this workshop, you should be able to….
- Base Principles:
  - Manipulate your data to extract the information you need for analysis in R
  - Generate a range of useful plots using R
- Presentation:
  - Customize the features of the plots and adapt them based on your needs
  - Understand principles of good design
- Exploration
  - Perform basic clustering and data reduction methods to visualize and identify relationships in your data
  - Generate interactive plots for easier data analysis

Why should we use R for data visualization?

- It offers a lot more control & flexibility over the plots & figures that you can generate
- CUSTOMIZATION
- More accurate representation of your data
- Increases reproducibility of your work

Presenting the datasets:

- Our datasets are called uvm_counts and uvm_clin
- Datasets can have whatever name you choose….
- With some caveats!
  - DON'T start the name with a number or a symbol

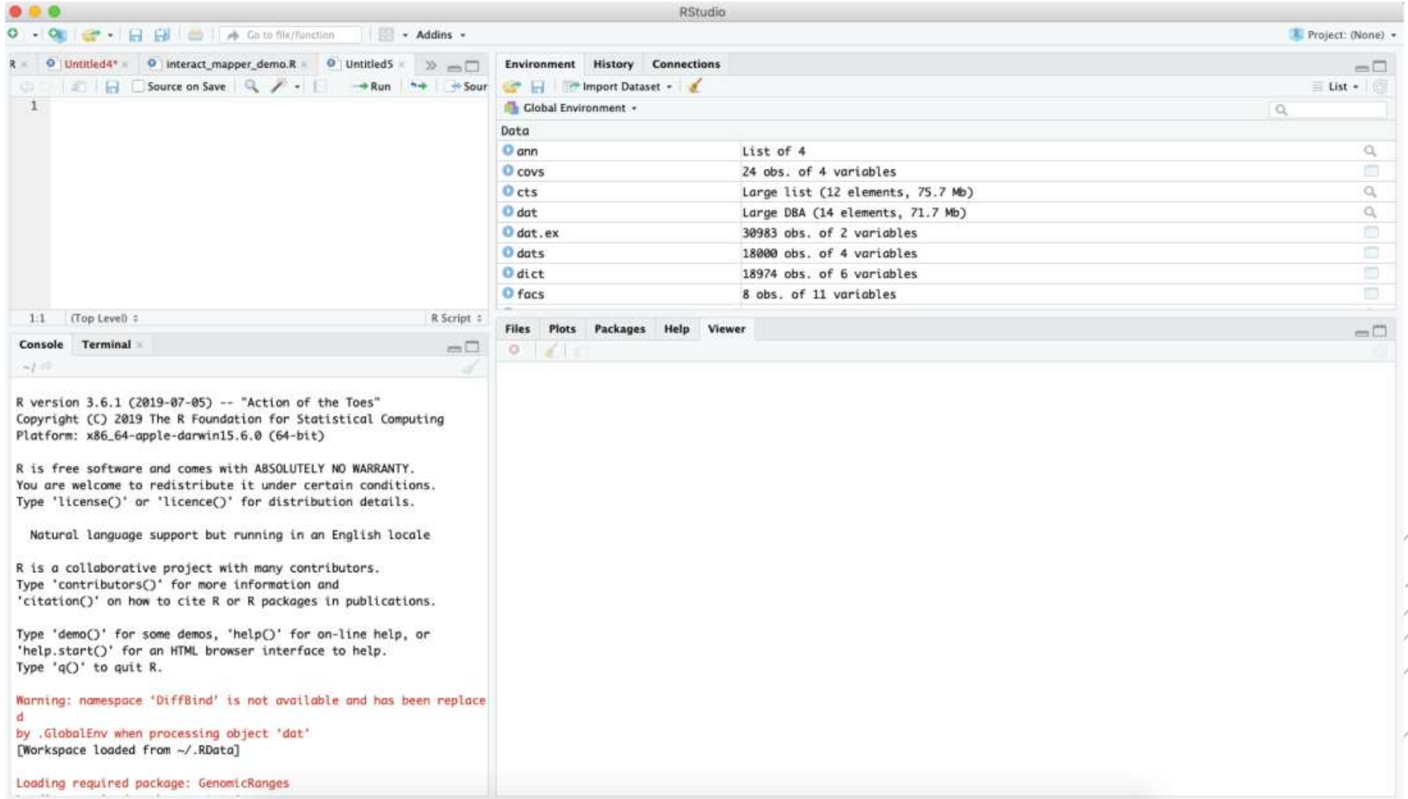
Context: UVM: Uveal Melanoma

- Rare disease compared to other cancers, but the most common eye cancer in adults
- Low local recurrence rate with treatment, but up to 50% of patients have metastasis
- No effective treatment for UVM metastasis currently available
- Survival time of metastatic patients less than 12 months after metastatic diagnosis

Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma, Robertson, A.G., 2017

RStudio Interface

- RStudio should look something like this!
- We're going to be typing up here to save our work and edit more easily.
- To run a command you typed, first highlight it with your cursor, then:
  - If you're on a Windows, press Ctrl and Enter
  - If you're on a Mac, press Command and enter
  - Voila!



Data	
ann	List of 4
covs	24 obs. of 4 variables
cts	Large list (12 elements, 75.7 Mb)
dat	Large DBA (14 elements, 71.7 Mb)
dat.ex	30983 obs. of 2 variables
dats	18000 obs. of 4 variables
dict	18974 obs. of 6 variables
facs	8 obs. of 11 variables

```
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Warning: namespace 'DiffBind' is not available and has been replace
d
by .GlobalEnv when processing object 'dat'
[Workspace loaded from ~/.RData]

Loading required package: GenomicRanges
```

Features of the Dataset

- **Let’s check out our data!**
- dim(tcga_express)

Range of rows you want to retain

- tcga_express[1:4,1:5]

Range of columns you want to retain

Structure of a ggplot command:

- Let’s start with a simple plot and build up from there!

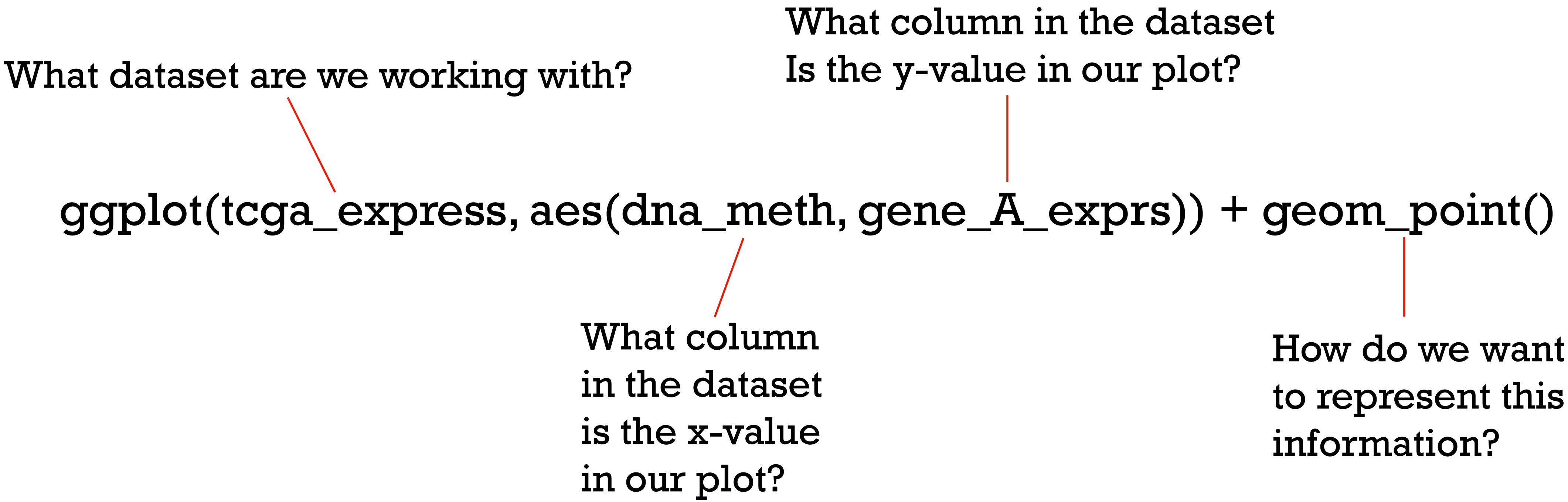
What dataset are we working with?

What column in the dataset
Is the y-value in our plot?

- ggplot(tcga_express, aes(x=dna_meth, y=gene_A_exprs)) + geom_point()

How do we want
to represent this
information?

What column in
the dataset
is the x-value in
our plot?



```
ggplot(tcga_express, aes(dna_meth, gene_A_exprs)) + geom_point()
```

What dataset are we working with?

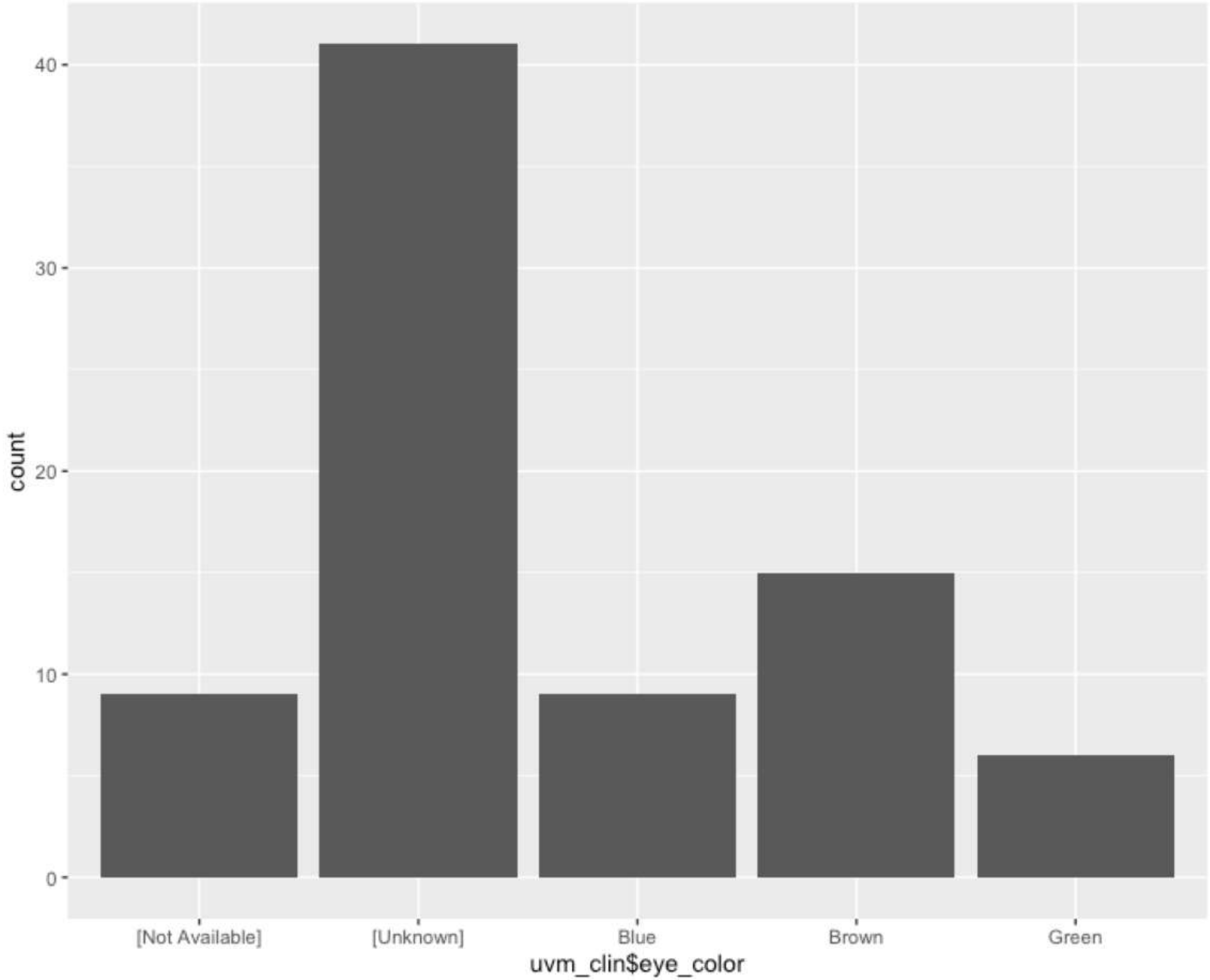
What column in the dataset
Is the y-value in our plot?

What column
in the dataset
is the x-value
in our plot?

How do we want
to represent this
information?

Barplot

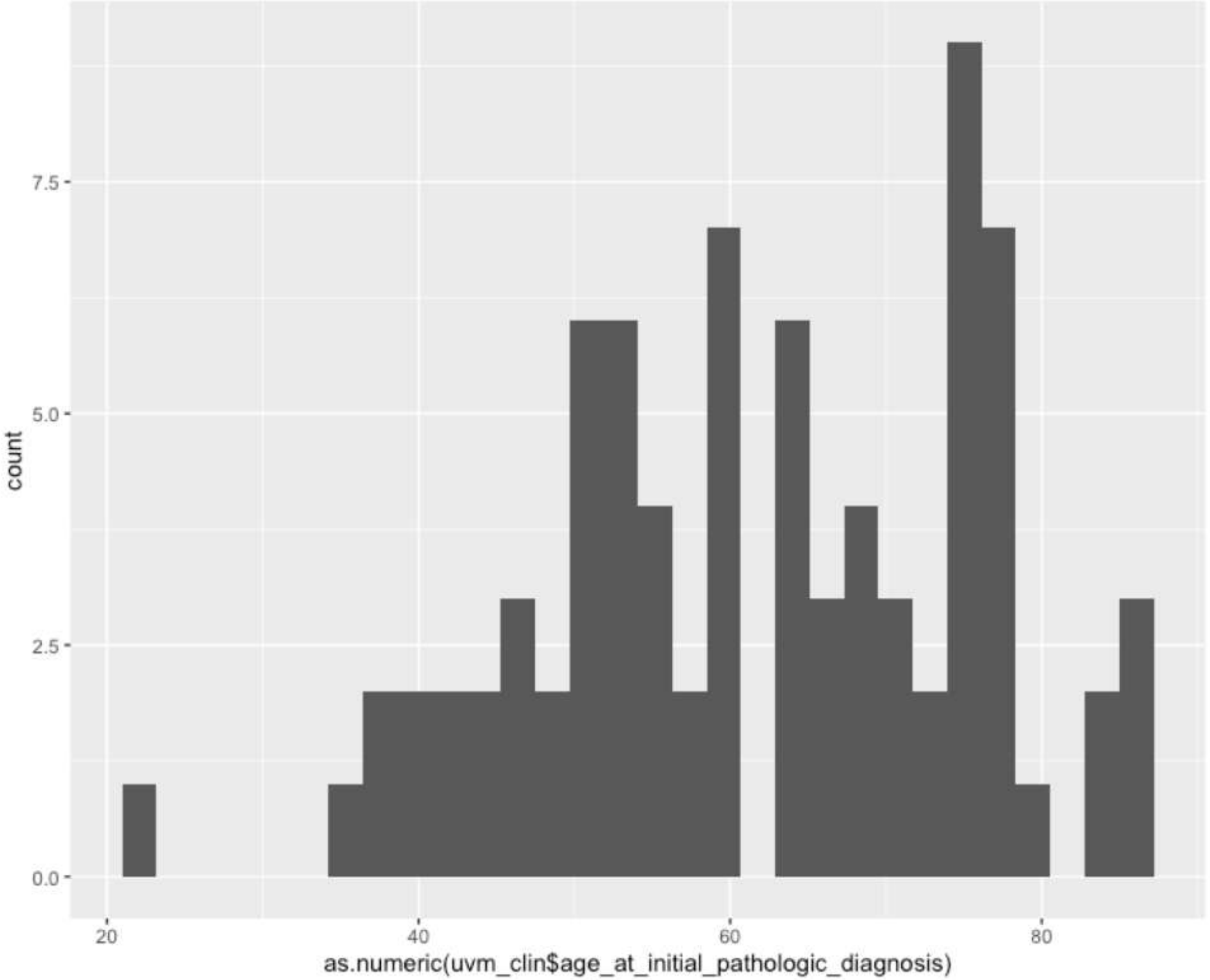
```
ggplot(uvm_clin, aes(uvm_clin$eye_color))+geom_bar()
```



Histogram

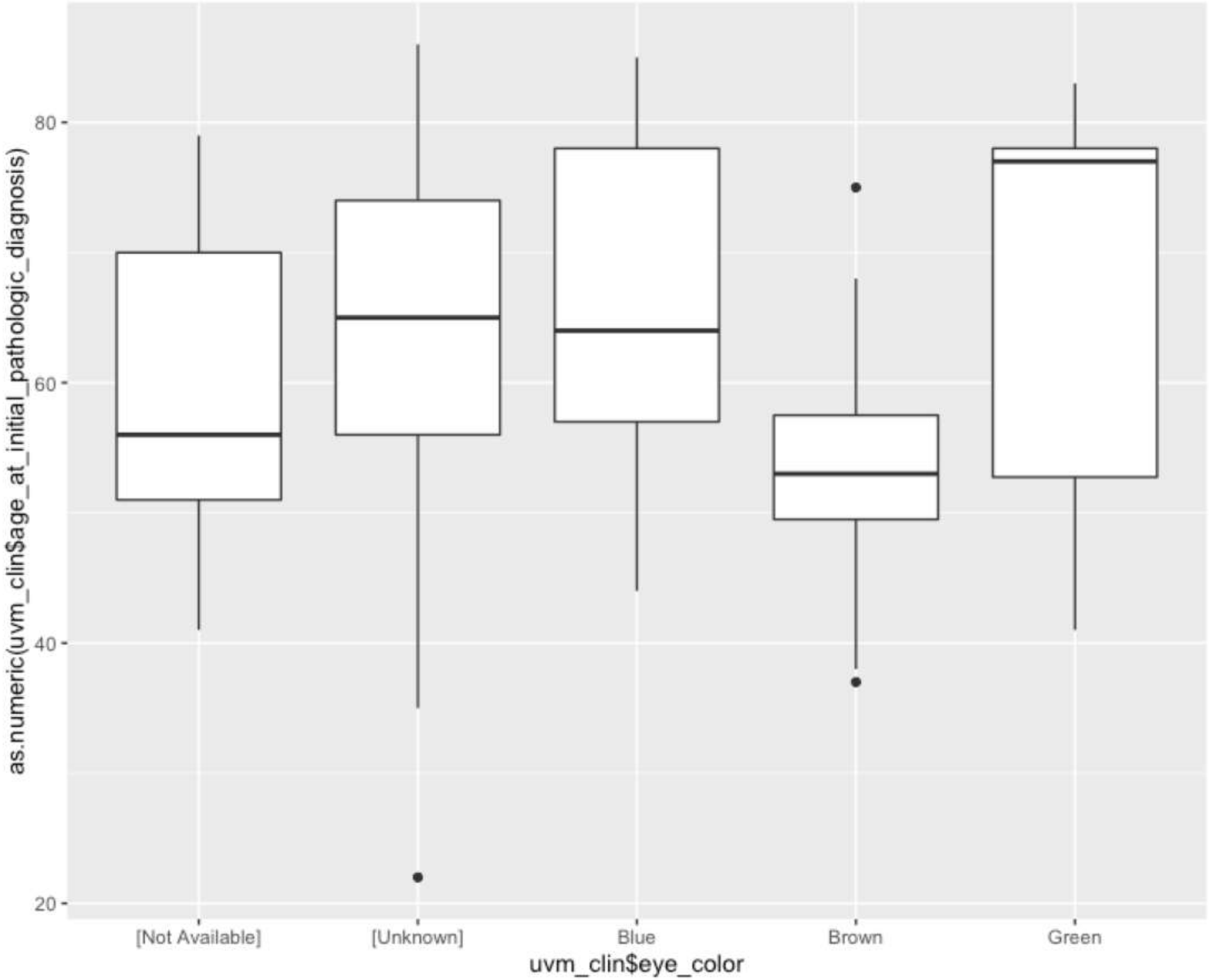
* Note that there is only one variable entered in the aes field!

```
ggplot(uvm_clin, aes(uvm_clin$age_at_initial_pathologic_diagnosis)) +
geom_histogram()
```



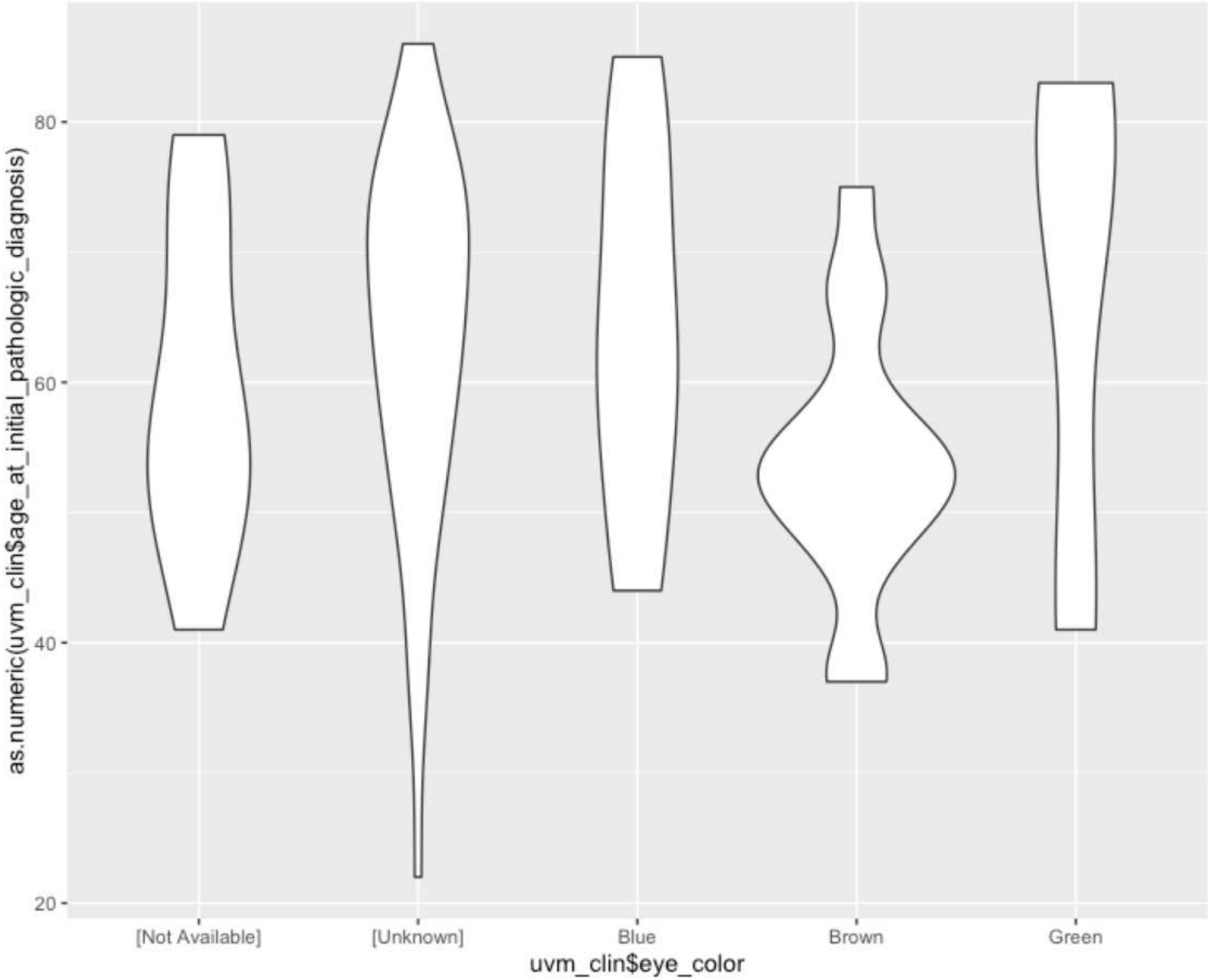
```
ggplot(uvm_clin, aes(x=uvm_clin$eye_color, y=uvm_clin$age_at_initial_pathologic_diagnosis))+geom_boxplot()
```

Boxplot



```
ggplot(uvm_clin, aes(x=uvm_clin$eye_color, y=uvm_clin$age_at_initial_pathologic_diagnosis))+geom_violin()
```

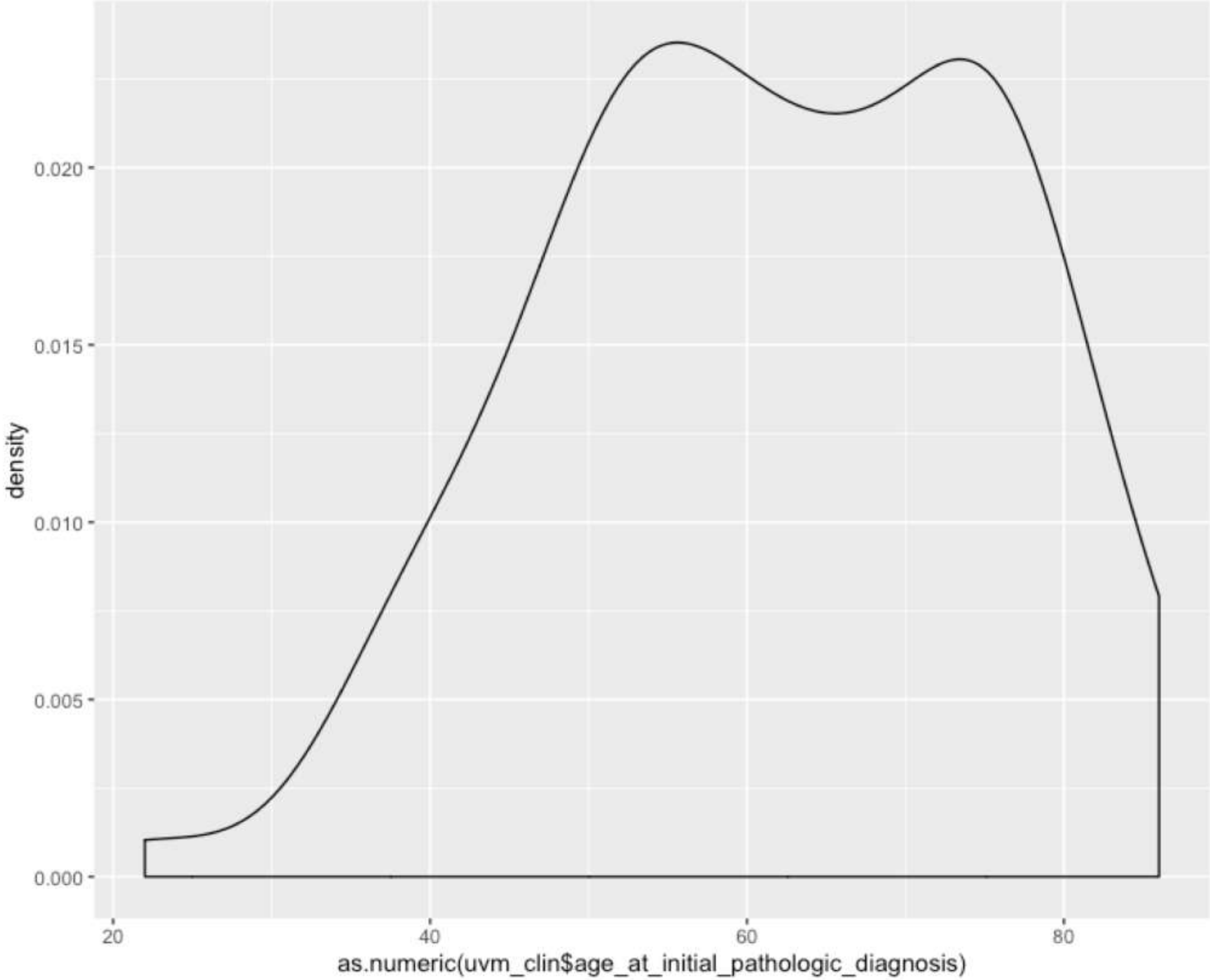
Violin Plot



Density Plot

* Note that there is only one variable entered in the aes field… for now!

```
ggplot(uvm_clin, aes(uvm_clin$age_at_initial_pathologic_diagnosis)) +
geom_density()
```



ggplot2 cheat sheet

```
ggplot(data = data_of_interest , aes( variables of interest)) +
```

What kind of plot do you want to make?

Type of plot	ggplot argument
Scatter plot	geom_point()
Histogram	geom_histogram()
Density plot	geom_density()
Bar Plot	geom_bar()
Violin Plot	geom_violin()
Box Plot	geom_boxplot()

How to export your figure to pdf

```
pdf(file="name_of_your_plot.pdf")
#make your plot here
ggplot(…)+…
dev.off()
```

Exercise 1:

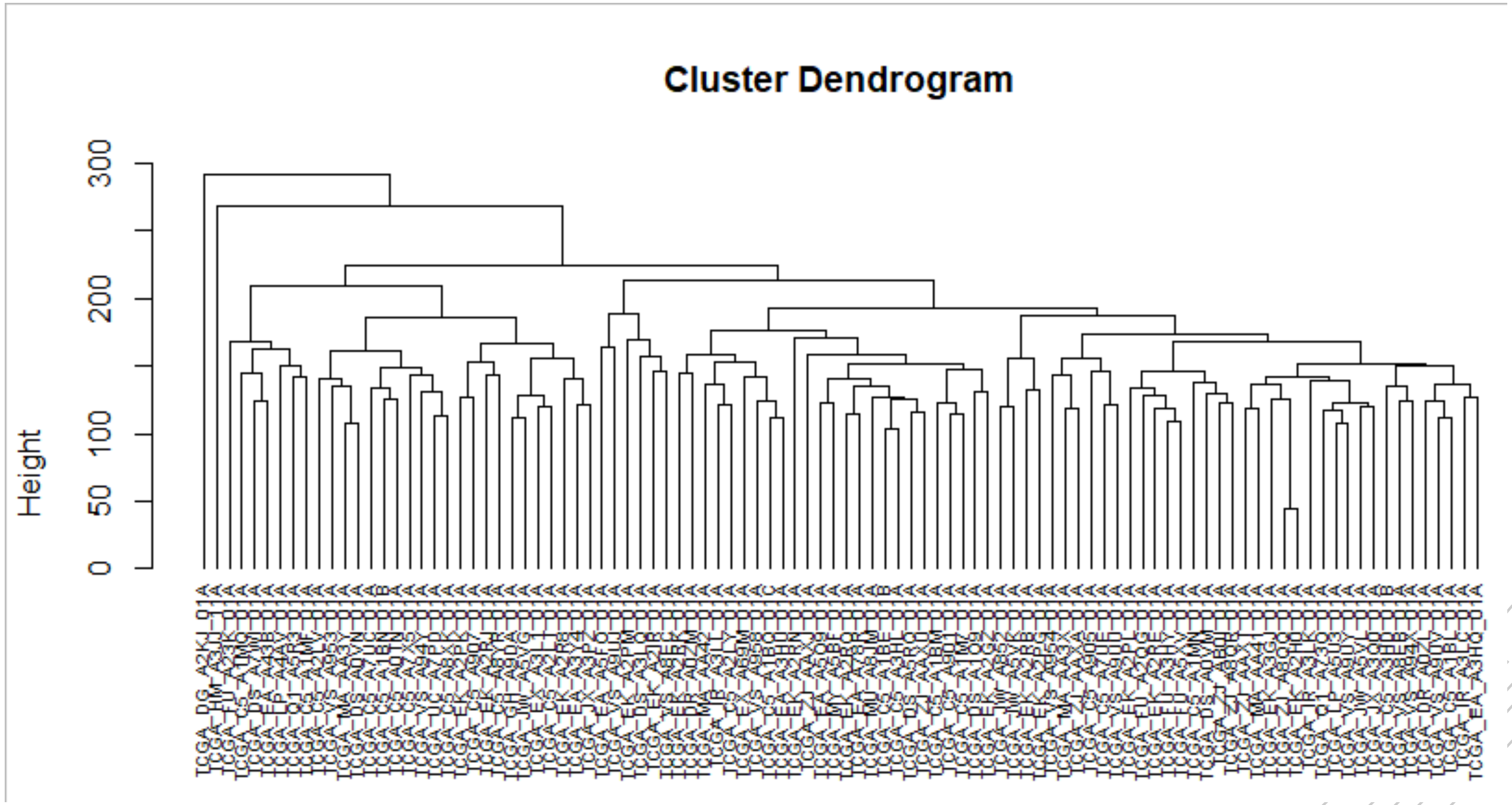
- You are trying to investigate the relationship between the expression of epigenetic modifier NSD1 and the expression of your favorite gene. Using the dataframe provided to you and your knowledge of ggplot2 commands so far, generate a plot illustrating this relationship.

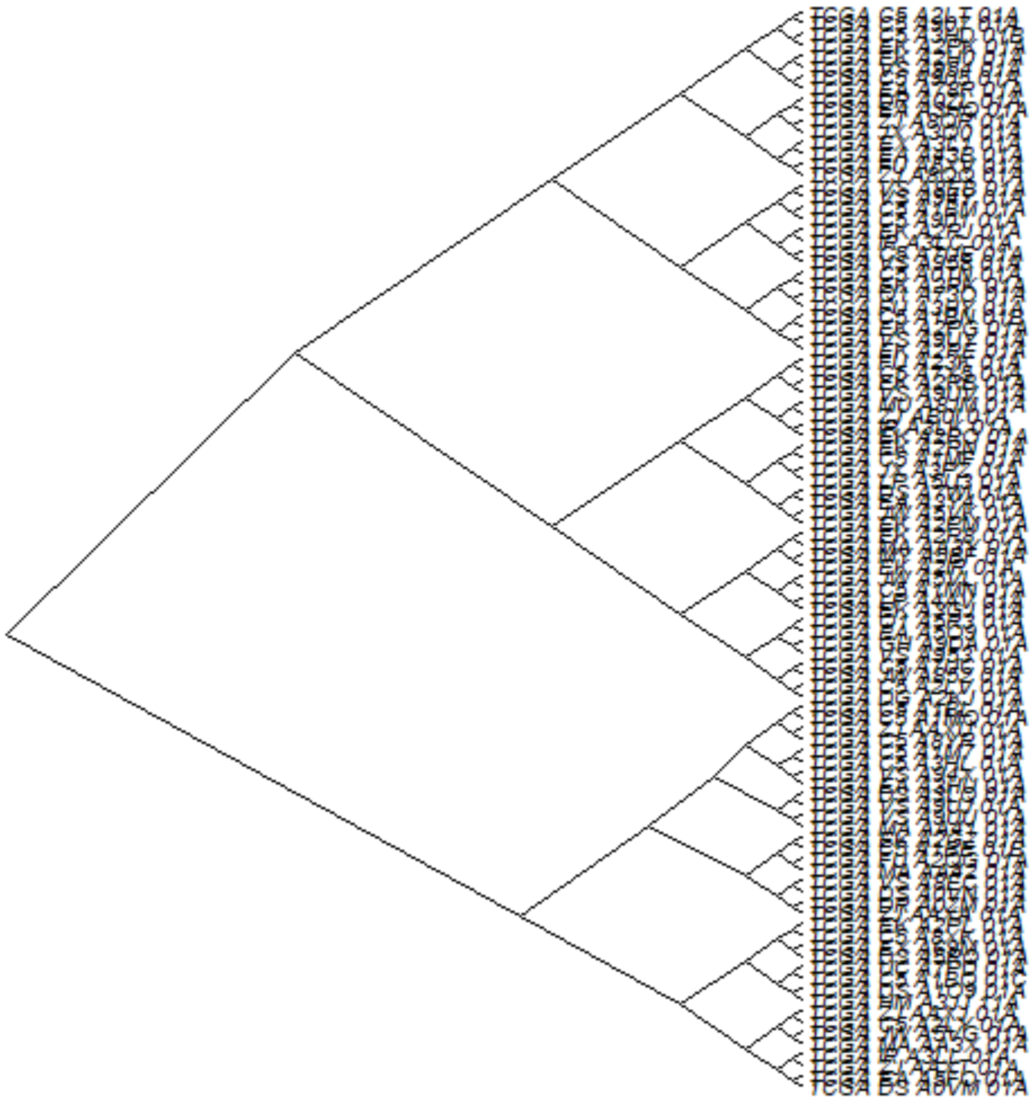
- Hint 1: start by breaking down the problem into manageable steps! Look at the data you have and the plot you want to generate. What steps do you need to do in order to bridge that gap?
- Hint 2: Here are some steps to help you on your way.
  - 1. Find out how to select only the information that you need from the data frame.
  - 2. Think of what plot is appropriate to represent this kind of data.
  - 3. What command in R will generate this plot?

Specialty Plots

Dendrogram

- clust_my_data<-hclust(dist(my_data))
- plot(as.dendrogram(clust_my_data))





Cladogram

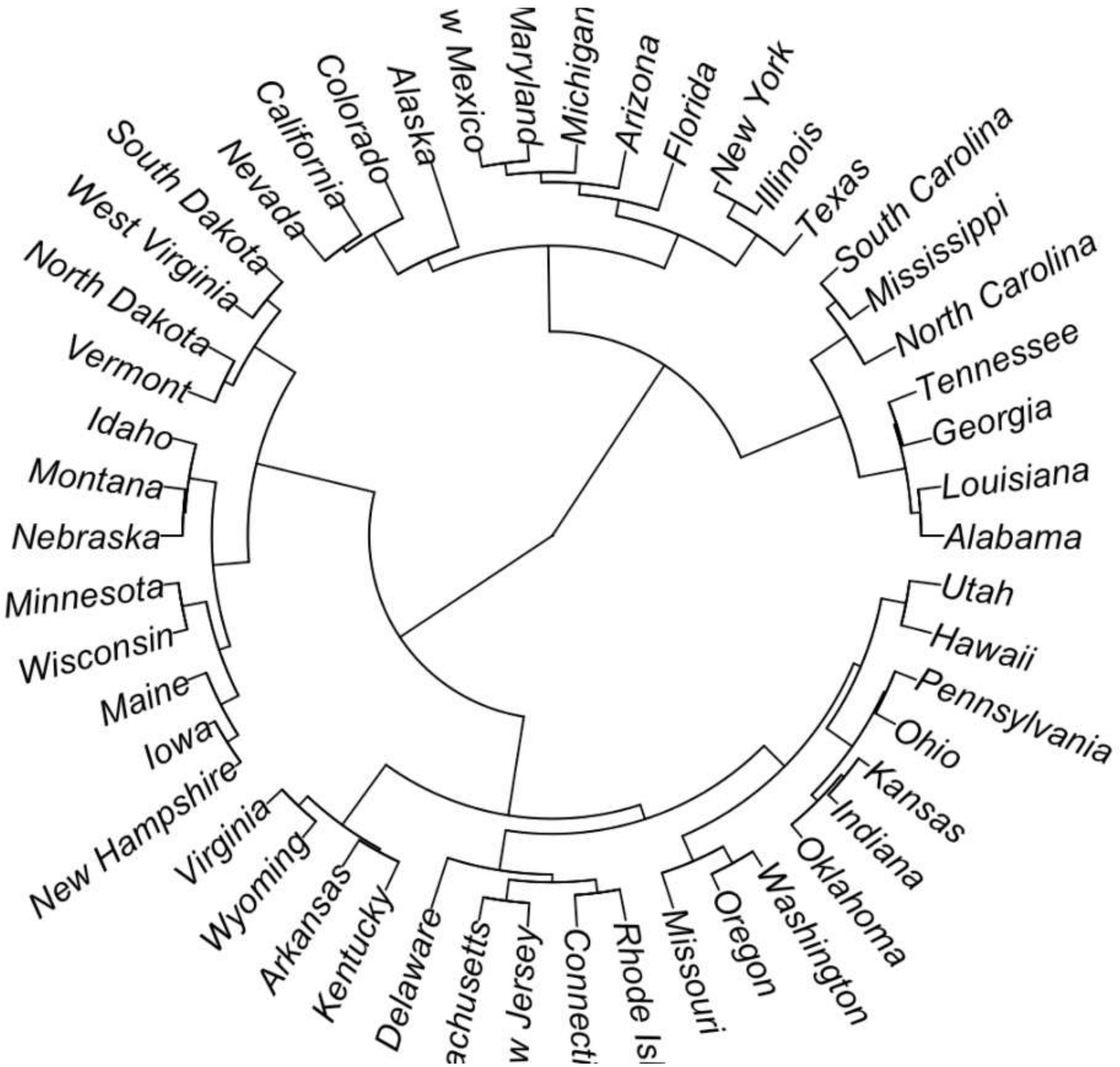
Install and load the ape package, then:

```
plot(as.phylo(clust_my_data), type = "cladogram")
```

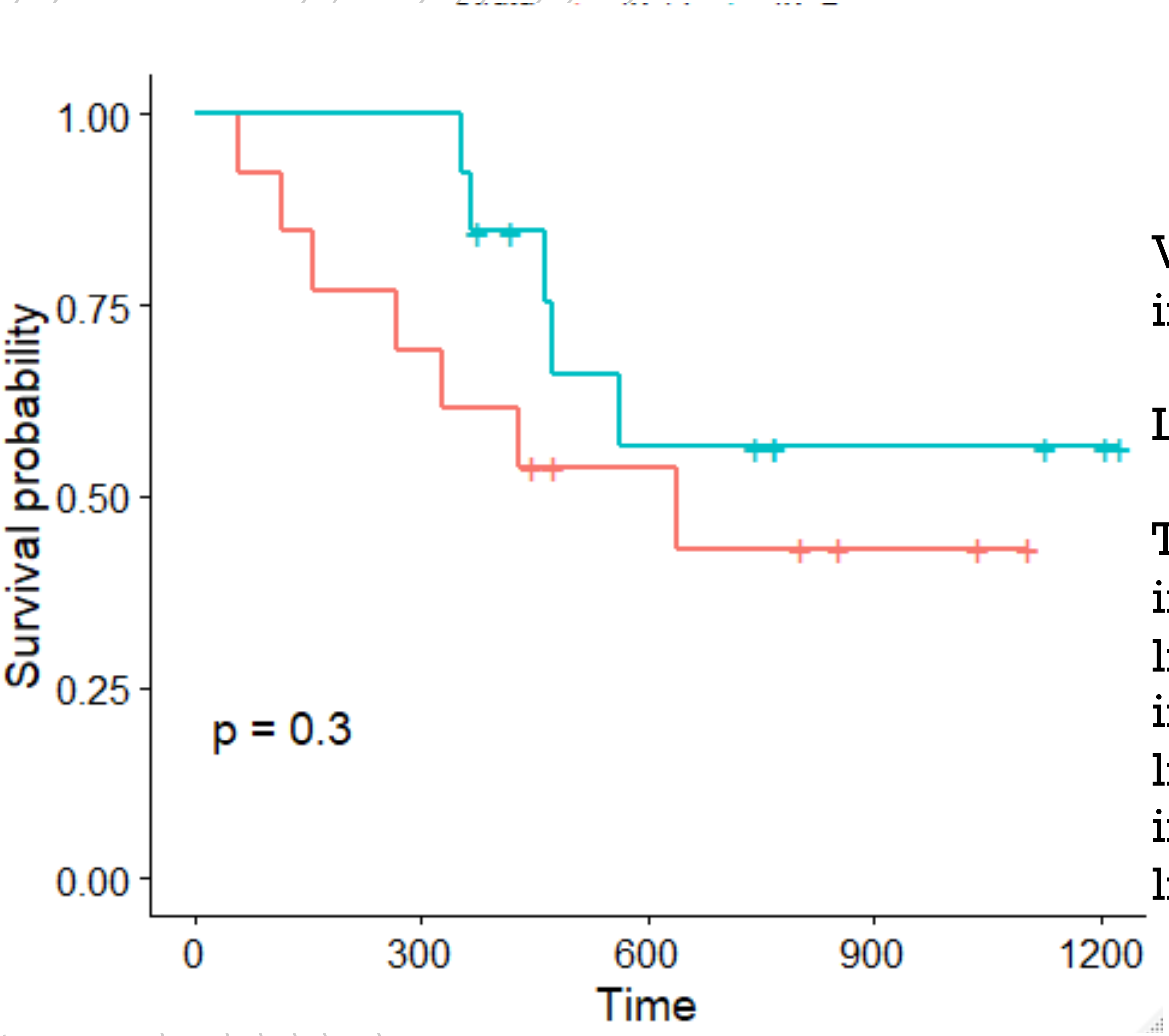
http://www.sthda.com/english/wiki/beautiful-dendrogram-visualizations-in-r-5-must-known-methods-unsupervised-machine-learning

How about a circular/fan dendrogram?

- plot(as.phylo(clust_my_data), type = "fan")



http://www.sthda.com/english/wiki/beautiful-dendrogram-visualizations-in-r-5-must-known-methods-unsupervised-machine-learning



Kaplan-Meier Curve

Very useful for assessing whether your feature of interest actually influences survival time or not!

Let’s give this a try!

To get started, install and load the following packages:

```
install.packages("survival”)
library(survival)
install.packages("survminer")
library(survminer)
install.packages("dplyr")
library(dplyr)
```

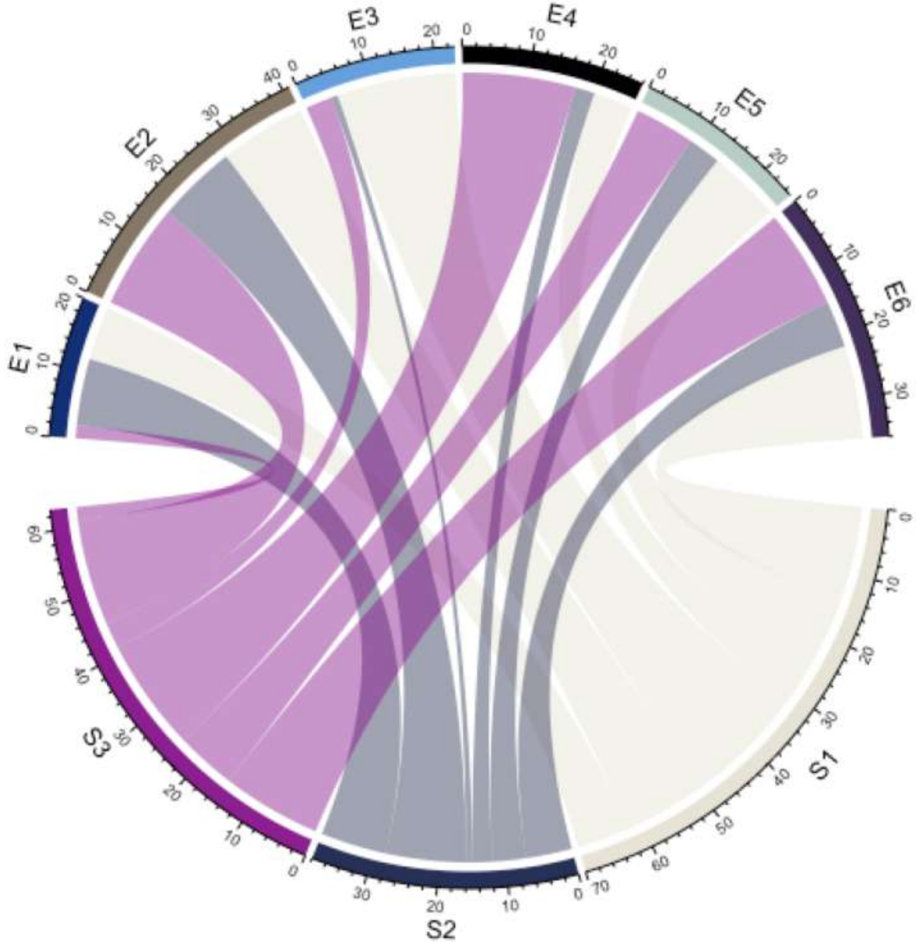
Ref: https://www.datacamp.com/community/tutorials/survival-analysis-R

Exercice 2

- Now that we made our first survival plot with the ovarian patient data, let’s apply this to our UVM dataset!
- Generate a survival plot for UVM patients using the feature of your choice (clinical feature, gene expression, etc)!

Chord Diagram Plot

Helpful for showing: chromosomal rearrangements, population migrations…



Elements of Figure Design!

- Colour
- Transparency
- Background
- Shape
- Size
- Trendlines & Regression lines

Questions to ask yourself when you're designing figures:

What do I look for in figures…

In a paper?

In a presentation?

How do these two answers differ?

Same data, different viewing experience, different design

Questions to ask yourself when you’re designing figures:

Who am I making this figure for/what is my audience?

- Myself/my lab
- Other scientists in my specialized field
- Other scientists (not necessarily in my field)
- Students
- General public for scientific communication

What are their needs?

Questions to ask yourself when you're designing figures:

Why am I making this figure?

NOT: I have to show this at lab meeting/I need material for my poster/Reviewer #2 asked about it

In the context of your work that you are presenting, what is the purpose of that figure?
Is it serving a role there or is it just filler?

What question am I answering in making it?

Is there a well-defined question that is behind the making of this figure, or is it just there to show work that was done?
Is it contributing to the reader's understanding?

What is the key message here?

If you are having troubles answering this, most likely your readers will too!!

Questions to ask yourself when you're designing figures:

Is my message clear and understandable?

Are the important features of the data appropriately highlighted?

Do your figures flow well and facilitate understanding?

Is my data easily seen and interpretable?

Is your data presented in a transparent, scientifically-responsible manner?

Does your figure encourage analysis and present opportunity for scientific dialogue, or does it obscure your data?

Questions to ask yourself when you’re designing figures:

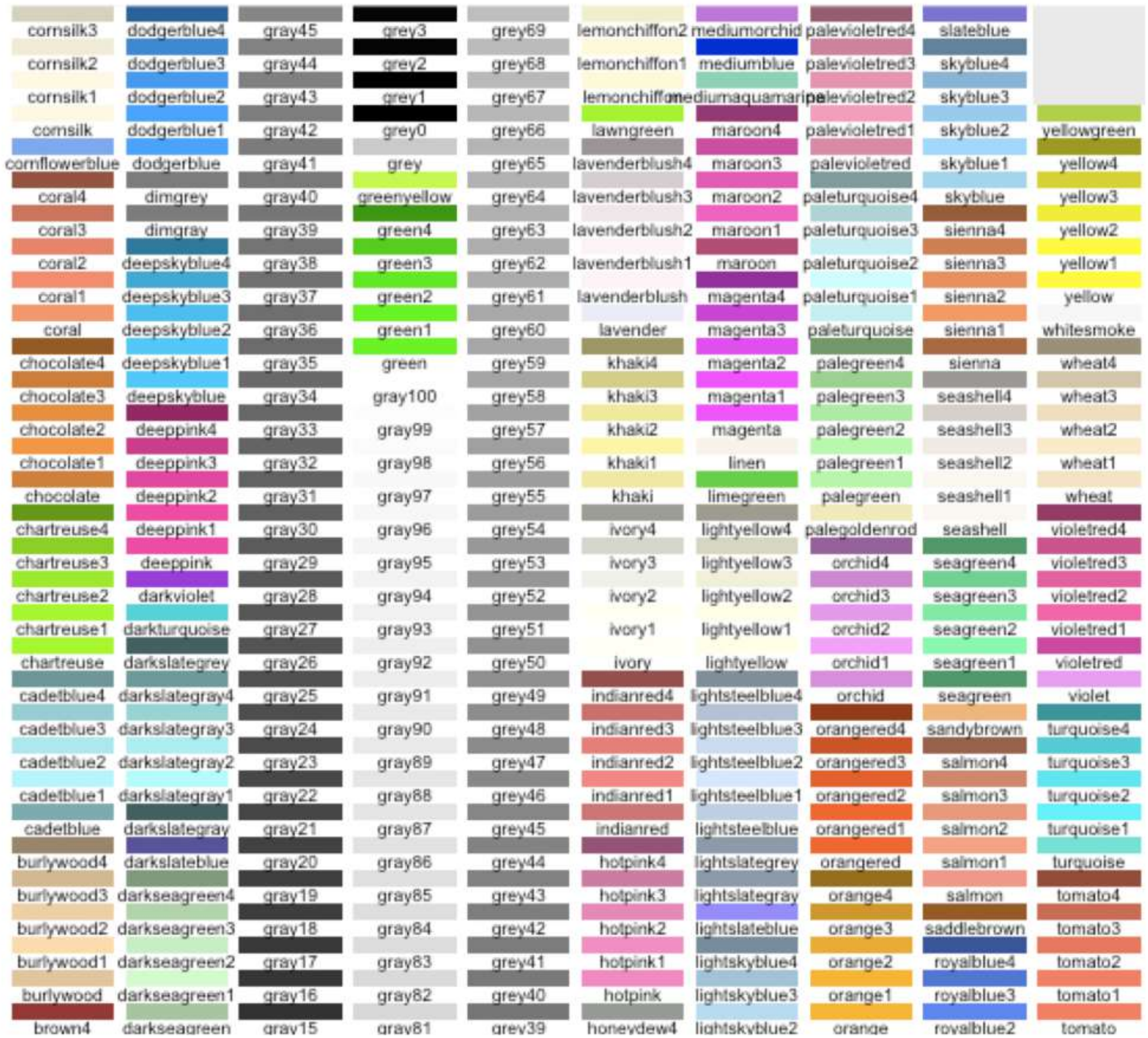
Is it visually appealing?

Why does this matter?

- Higher quality publications
- More successful posters and presentations
- Better scientific communication
- Higher engagement and interest your research

Colour

- Colour is a key aspect to making sure that your figures are:
- 1) conveying a **clear, strong message;**
- 2) **easily understood;**
- 3) **aesthetically pleasing.**



cornsilk3
cornsilk2
cornsilk1
cornsilk
cornflowerblue
coral4
coral3
coral2
coral1
coral
chocolate4
chocolate3
chocolate2
chocolate1
chocolate
chartreuse4
chartreuse3
chartreuse2
chartreuse1
chartreuse
cadetblue4
cadetblue3
cadetblue2
cadetblue1
cadetblue
burlywood4
burlywood3
burlywood2
burlywood1
burlywood
brown4

dodgerblue4
dodgerblue3
dodgerblue2
dodgerblue1
dodgerblue
dimgrey
dimgray
deepskyblue4
deepskyblue3
deepskyblue2
deepskyblue1
deepskyblue
deeppink4
deeppink3
deeppink2
deeppink1
deeppink
darkviolet
darkturquoise
darkslategrey
darkslategray4
darkslategray3
darkslategray2
darkslategray1
darkslategray
darkslateblue
darkseagreen4
darkseagreen3
darkseagreen2
darkseagreen1
darkseagreen

gray45
gray44
gray43
gray42
gray41
gray40
gray39
gray38
gray37
gray36
gray35
gray34
gray33
gray32
gray31
gray30
gray29
gray28
gray27
gray26
gray25
gray24
gray23
gray22
gray21
gray20
gray19
gray18
gray17
gray16
gray15

grey3
grey2
grey1
grey0
grey
greenyellow
green4
green3
green2
green1
green
gray100
gray99
gray98
gray97
gray96
gray95
gray94
gray93
gray92
gray91
gray90
gray89
gray88
gray87
gray86
gray85
gray84
gray83
gray82
gray81

grey69
grey68
grey67
grey66
grey65
grey64
grey63
grey62
grey61
grey60
grey59
grey58
grey57
grey56
grey55
grey54
grey53
grey52
grey51
grey50
grey49
grey48
grey47
grey46
grey45
grey44
grey43
grey42
grey41
grey40
grey39

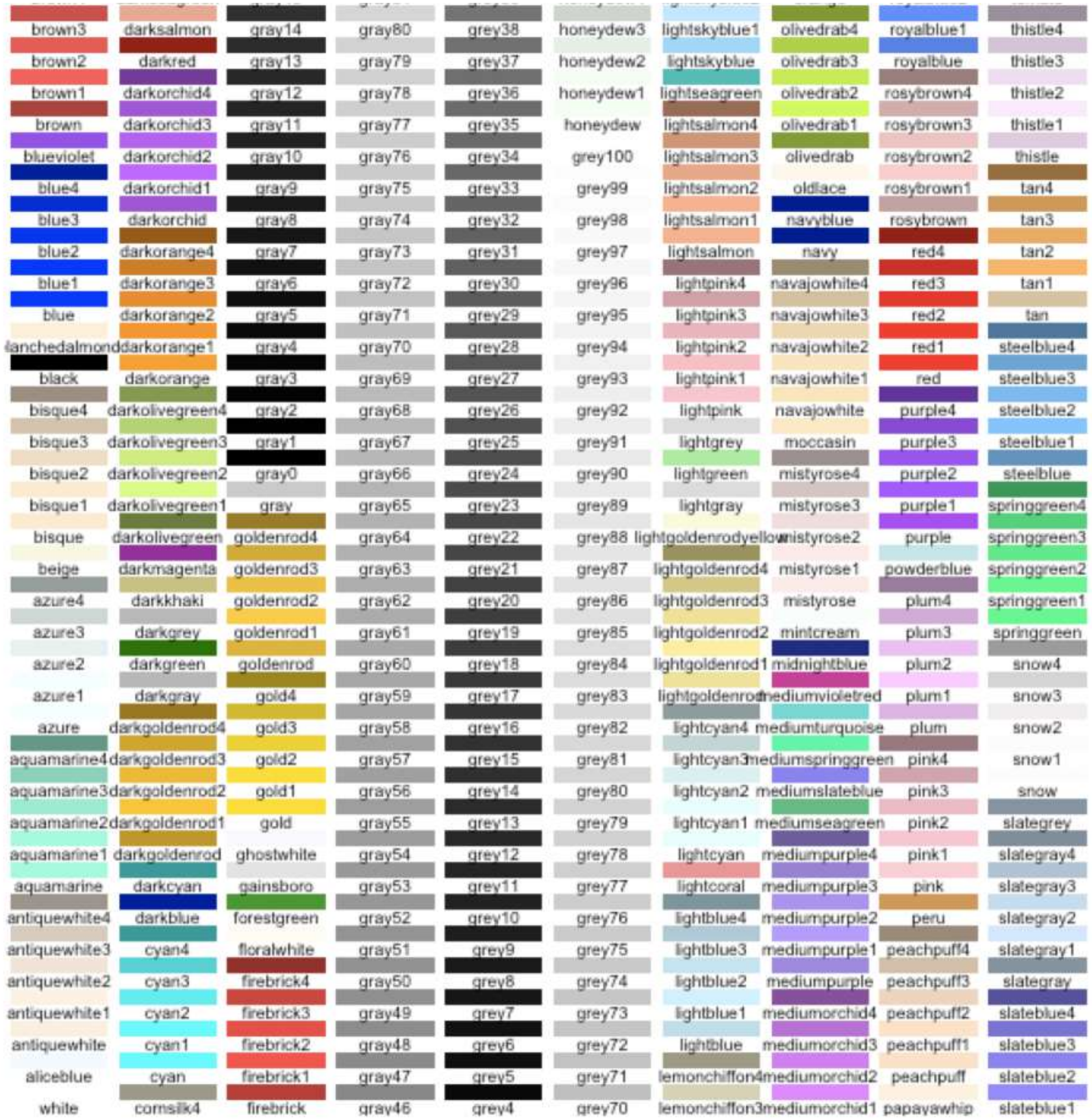
lemonchiffon2
lemonchiffon1
lemonchiffon
lawngreen
lavenderblush4
lavenderblush3
lavenderblush2
lavenderblush1
lavenderblush
lavender
khaki4
khaki3
khaki2
khaki1
khaki
ivory4
ivory3
ivory2
ivory1
ivory
indianred4
indianred3
indianred2
indianred1
indianred
hotpink4
hotpink3
hotpink2
hotpink1
hotpink
honeydew4

mediumorchid
mediumblue
mediumaquamarine
maroon4
maroon3
maroon2
maroon1
maroon
magenta4
magenta3
magenta2
magenta1
magenta
linen
limegreen
lightyellow4
lightyellow3
lightyellow2
lightyellow1
lightyellow
lightsteelblue4
lightsteelblue3
lightsteelblue2
lightsteelblue1
lightsteelblue
lightslategrey
lightslategray
lightslateblue
lightskyblue4
lightskyblue3
lightskyblue2

palevioletred4
palevioletred3
palevioletred2
palevioletred1
palevioletred
paleturquoise4
paleturquoise3
paleturquoise2
paleturquoise1
paleturquoise
palegreen4
palegreen3
palegreen2
palegreen1
palegreen
palegoldenrod
orchid4
orchid3
orchid2
orchid1
orchid
orangered4
orangered3
orangered2
orangered1
orangered
orange4
orange3
orange2
orange1
orange

slateblue
skyblue4
skyblue3
skyblue2
skyblue1
skyblue
sienna4
sienna3
sienna2
sienna1
sienna
seashell4
seashell3
seashell2
seashell1
seashell
seagreen4
seagreen3
seagreen2
seagreen1
seagreen
sandybrown
salmon4
salmon3
salmon2
salmon1
salmon
saddlebrown
royalblue4
royalblue3
royalblue2

yellowgreen
yellow4
yellow3
yellow2
yellow1
yellow
whitesmoke
wheat4
wheat3
wheat2
wheat1
wheat
violetred4
violetred3
violetred2
violetred1
violetred
violet
turquoise4
turquoise3
turquoise2
turquoise1
turquoise
tomato4
tomato3
tomato2
tomato1
tomato



brown3
brown2
brown1
brown
blueviolet
blue4
blue3
blue2
blue1
blue
lanchedalmond
black
bisque4
bisque3
bisque2
bisque1
bisque
beige
azure4
azure3
azure2
azure1
azure
aquamarine4
aquamarine3
aquamarine2
aquamarine1
aquamarine
antiquewhite4
antiquewhite3
antiquewhite2
antiquewhite1
antiquewhite
aliceblue
white

darksalmon
darkred
darkorchid4
darkorchid3
darkorchid2
darkorchid1
darkorchid
darkorange4
darkorange3
darkorange2
darkorange1
darkorange
darkolivegreen4
darkolivegreen3
darkolivegreen2
darkolivegreen1
darkolivegreen
darkmagenta
darkkhaki
darkgrey
darkgreen
darkgray
darkgoldenrod4
darkgoldenrod3
darkgoldenrod2
darkgoldenrod1
darkgoldenrod
darkcyan
darkblue
cyan4
cyan3
cyan2
cyan1
cyan
cornsilk4

gray14
gray13
gray12
gray11
gray10
gray9
gray8
gray7
gray6
gray5
gray4
gray3
gray2
gray1
gray0
gray
goldenrod4
goldenrod3
goldenrod2
goldenrod1
goldenrod
gold4
gold3
gold2
gold1
gold
ghostwhite
gainsboro
forestgreen
floralwhite
firebrick4
firebrick3
firebrick2
firebrick1
firebrick

gray80
gray79
gray78
gray77
gray76
gray75
gray74
gray73
gray72
gray71
gray70
gray69
gray68
gray67
gray66
gray65
gray64
gray63
gray62
gray61
gray60
gray59
gray58
gray57
gray56
gray55
gray54
gray53
gray52
gray51
gray50
gray49
gray48
gray47
gray46

grey38
grey37
grey36
grey35
grey34
grey33
grey32
grey31
grey30
grey29
grey28
grey27
grey26
grey25
grey24
grey23
grey22
grey21
grey20
grey19
grey18
grey17
grey16
grey15
grey14
grey13
grey12
grey11
grey10
grey9
grey8
grey7
grey6
grey5
grey4

honeydew3
honeydew2
honeydew1
honeydew
grey100
grey99
grey98
grey97
grey96
grey95
grey94
grey93
grey92
grey91
grey90
grey89
grey88
grey87
grey86
grey85
grey84
grey83
grey82
grey81
grey80
grey79
grey78
grey77
grey76
grey75
grey74
grey73
grey72
grey71
grey70

lightskyblue1
lightskyblue
lightseagreen
lightsalmon4
lightsalmon3
lightsalmon2
lightsalmon1
lightsalmon
lightpink4
lightpink3
lightpink2
lightpink1
lightpink
lightgrey
lightgreen
lightgray
lightgoldenrodyellow
lightgoldenrod4
lightgoldenrod3
lightgoldenrod2
lightgoldenrod1
lightgoldenrod
lightcyan4
lightcyan3
lightcyan2
lightcyan1
lightcyan
lightcoral
lightblue4
lightblue3
lightblue2
lightblue1
lightblue
lemonchiffon4
lemonchiffon3

olivedrab4
olivedrab3
olivedrab2
olivedrab1
olivedrab
oldlace
navyblue
navy
navajowhite4
navajowhite3
navajowhite2
navajowhite1
navajowhite
moccasin
mistyrose4
mistyrose3
mistyrose2
mistyrose1
mistyrose
mintcream
midnightblue
mediumvioletred
mediumturquoise
mediumspringgreen
mediumslateblue
mediumseagreen
mediumpurple4
mediumpurple3
mediumpurple2
mediumpurple1
mediumpurple
mediumorchid4
mediumorchid3
mediumorchid2
mediumorchid1

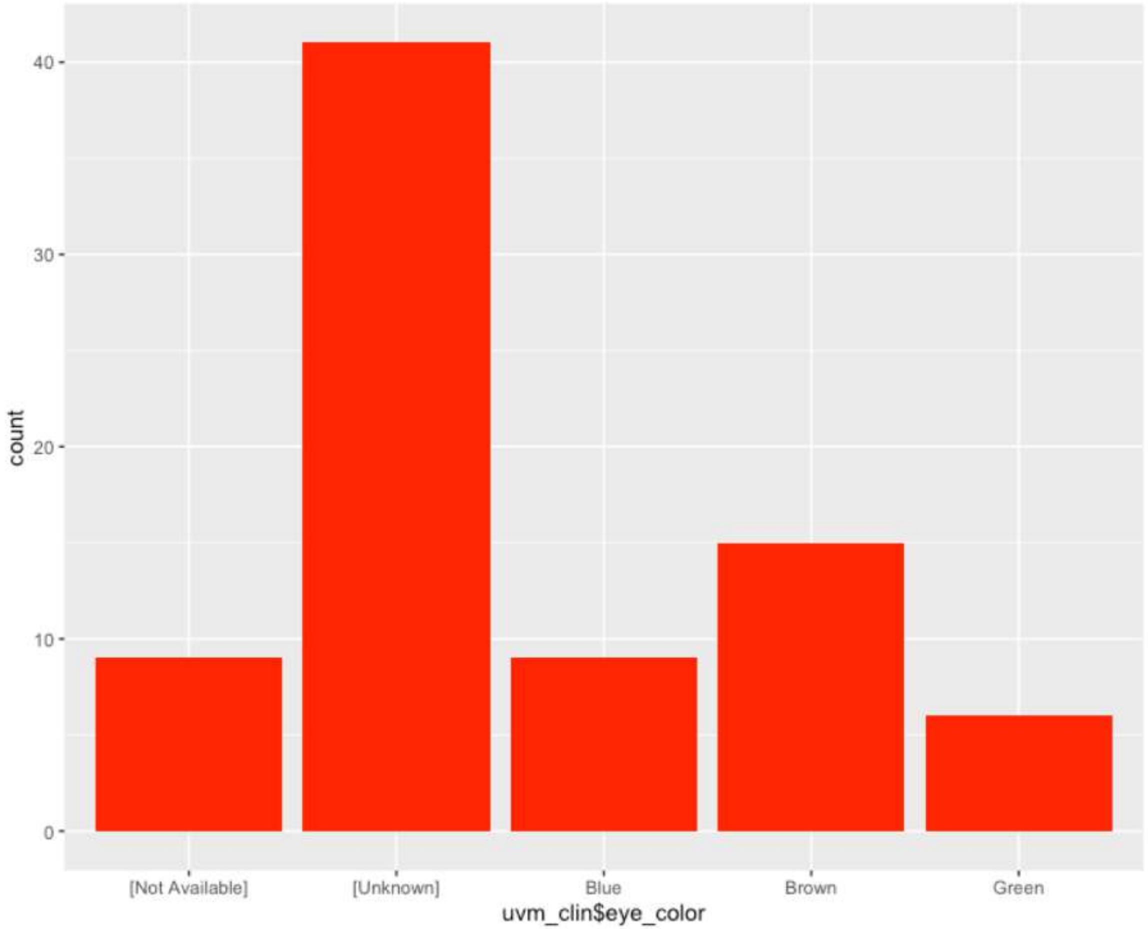
royalblue1
royalblue
rosybrown4
rosybrown3
rosybrown2
rosybrown1
rosybrown
red4
red3
red2
red1
red
purple4
purple3
purple2
purple1
purple
powderblue
plum4
plum3
plum2
plum1
plum
pink4
pink3
pink2
pink1
pink
peru
peachpuff4
peachpuff3
peachpuff2
peachpuff1
peachpuff
papayawhip

thistle4
thistle3
thistle2
thistle1
thistle
tan4
tan3
tan2
tan1
tan
steelblue4
steelblue3
steelblue2
steelblue1
steelblue
springgreen4
springgreen3
springgreen2
springgreen1
springgreen
snow4
snow3
snow2
snow1
snow
slategrey
slategray4
slategray3
slategray2
slategray1
slategray
slateblue4
slateblue3
slateblue2
slateblue1

Color: Uniform

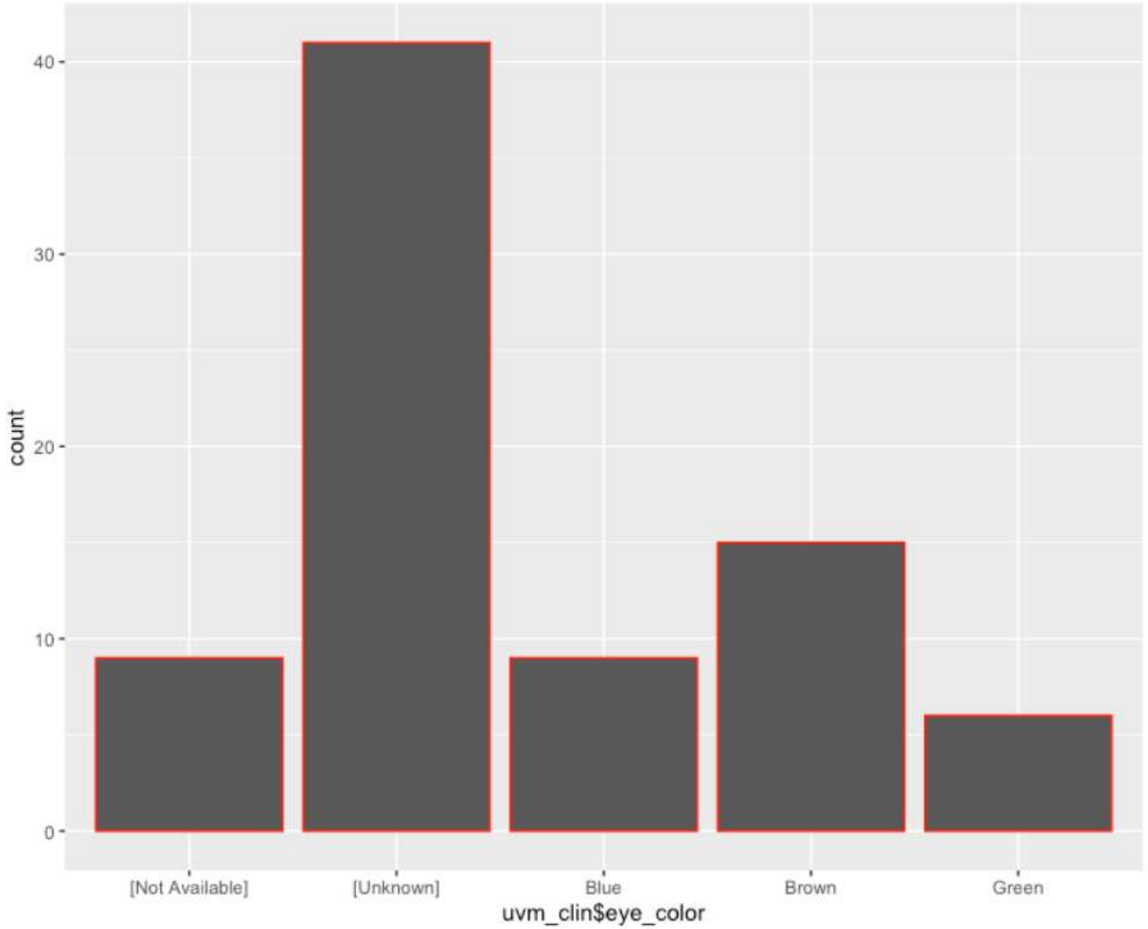
To change the bar colour:

```
geom_bar(fill = "red")
```



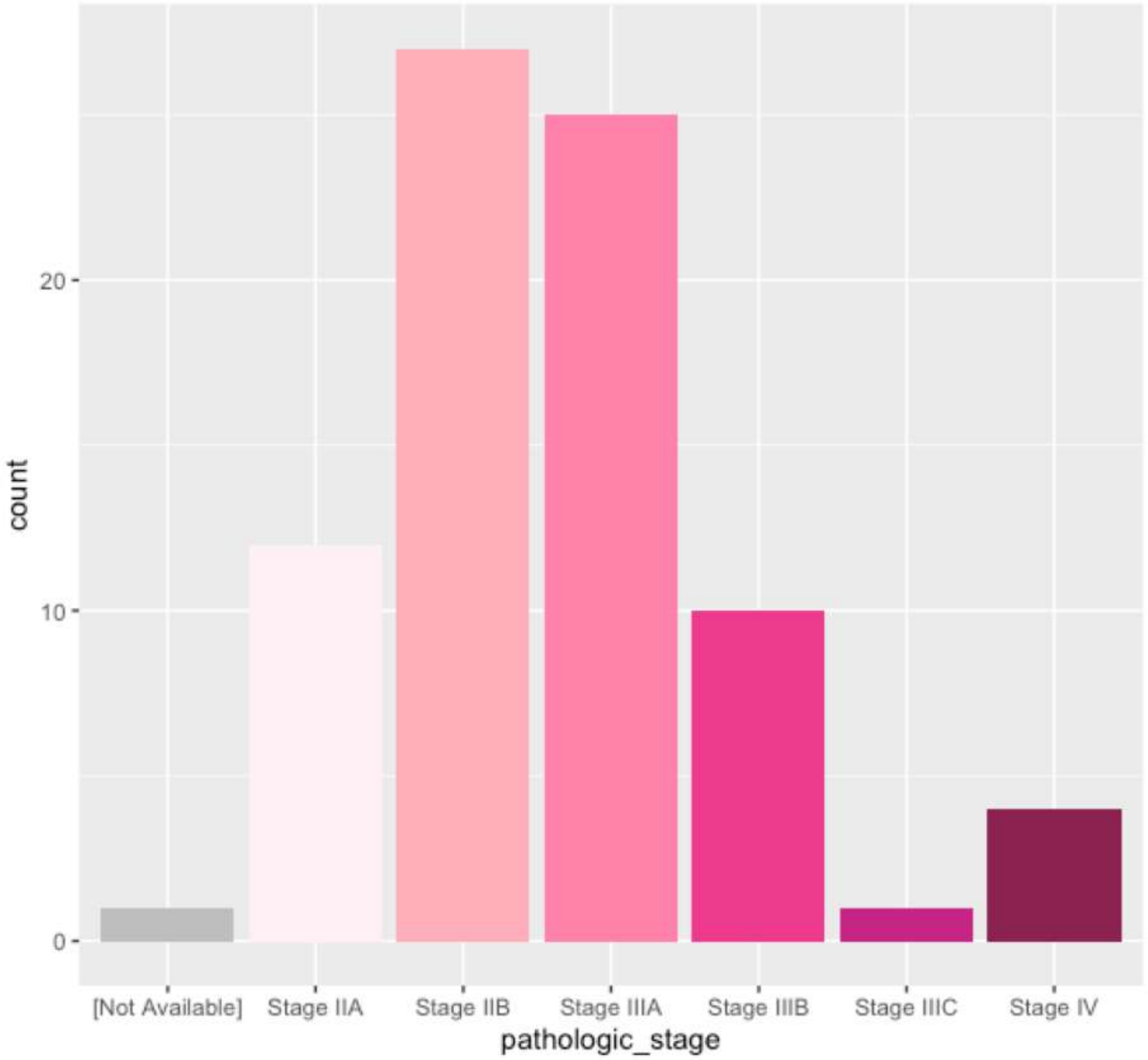
To change the colour of the bar’s outline:

```
+ geom_bar(colour = "red")
```

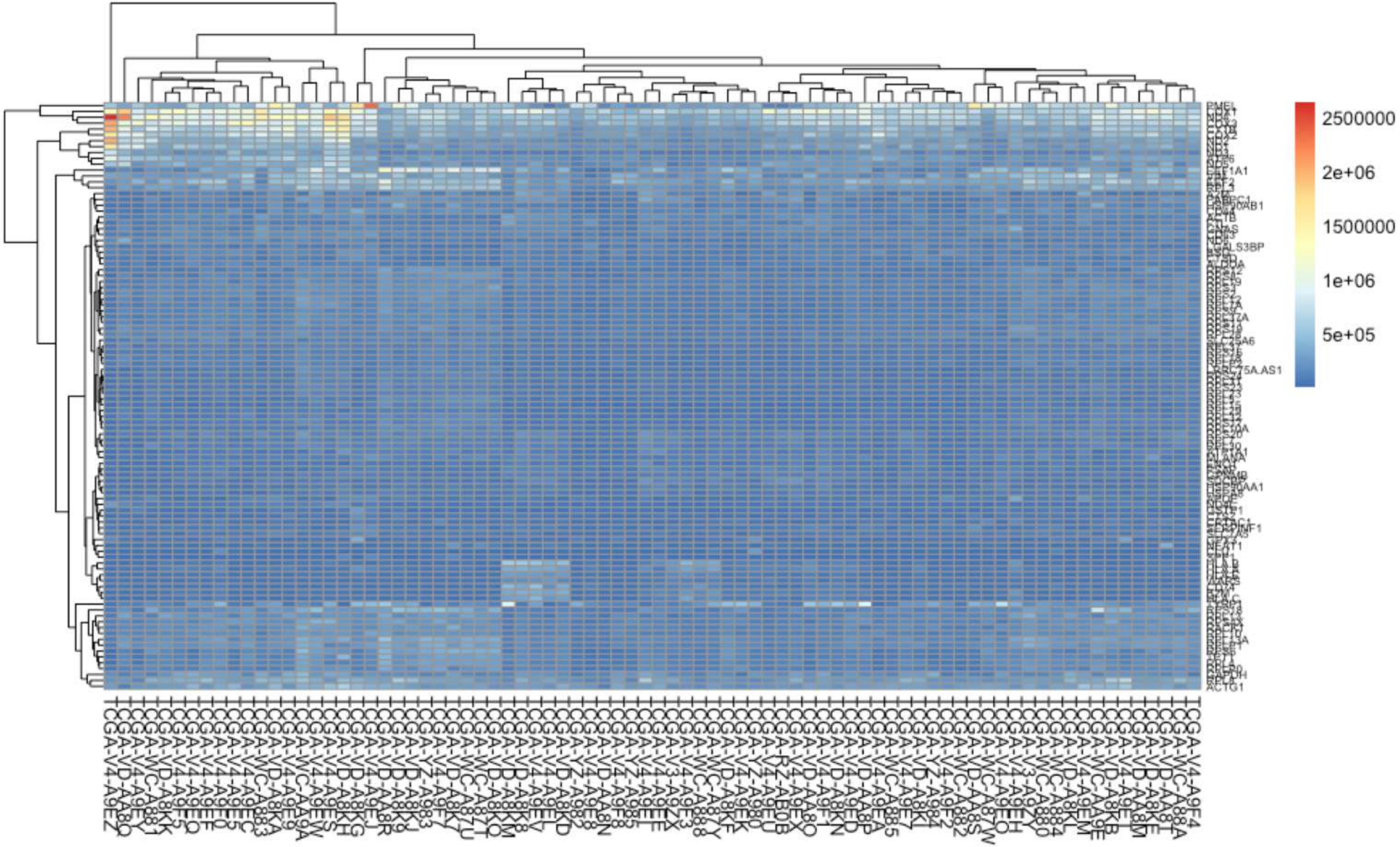


Color: Gradient

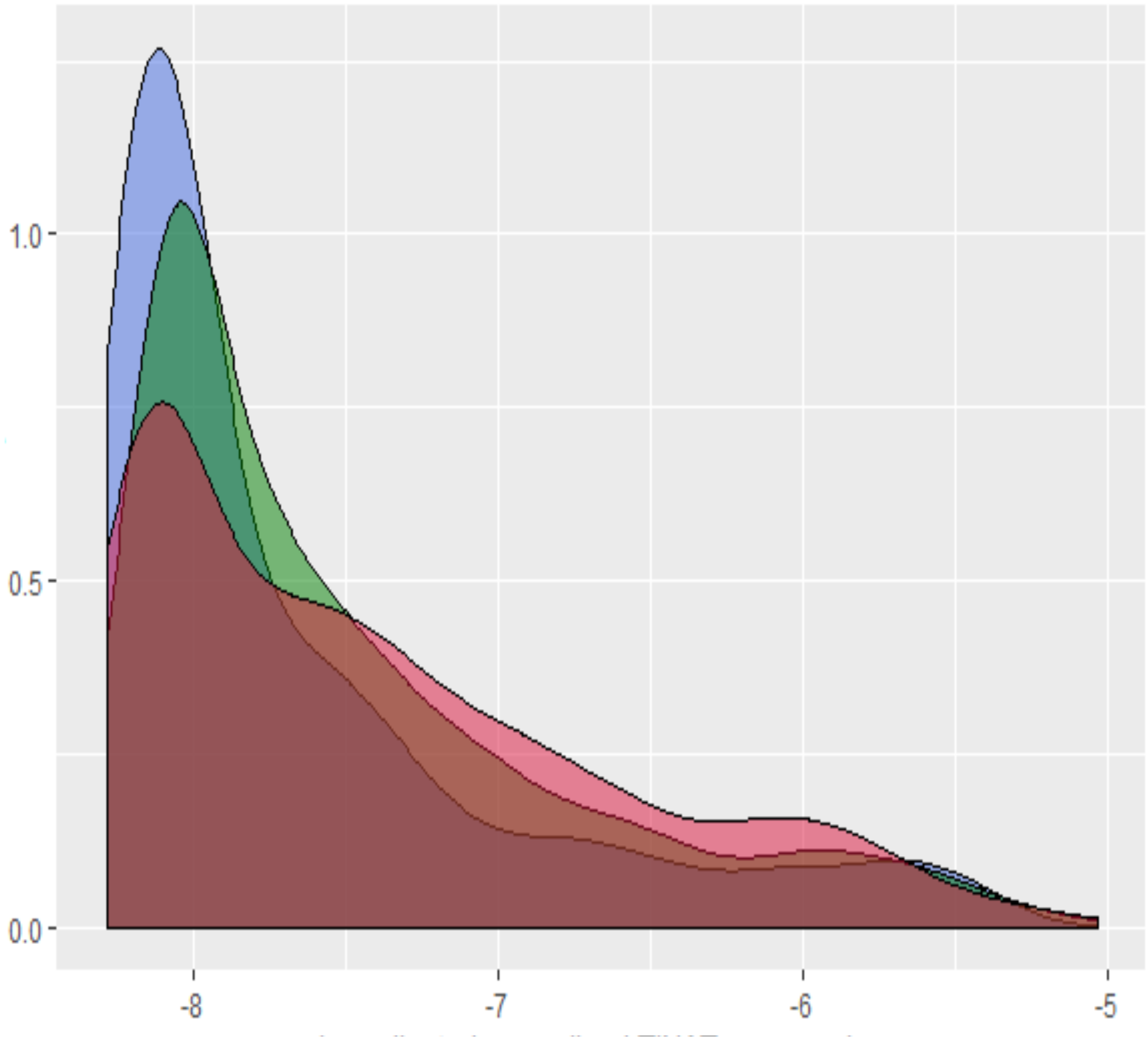
```
ggplot(uvm_clin, aes(pathologic_stage)) +
geom_bar(fill=c("grey", "lavenderblush1", "lightpink1", "palevioletred1",
"violetred2", "mediumvioletred", "violetred4"))
```



Color: Diverging



Color: Categorical



Selecting Your Colour Palettes:

Make it intuitive and know your field!

DNA methylation: Blue to white to red

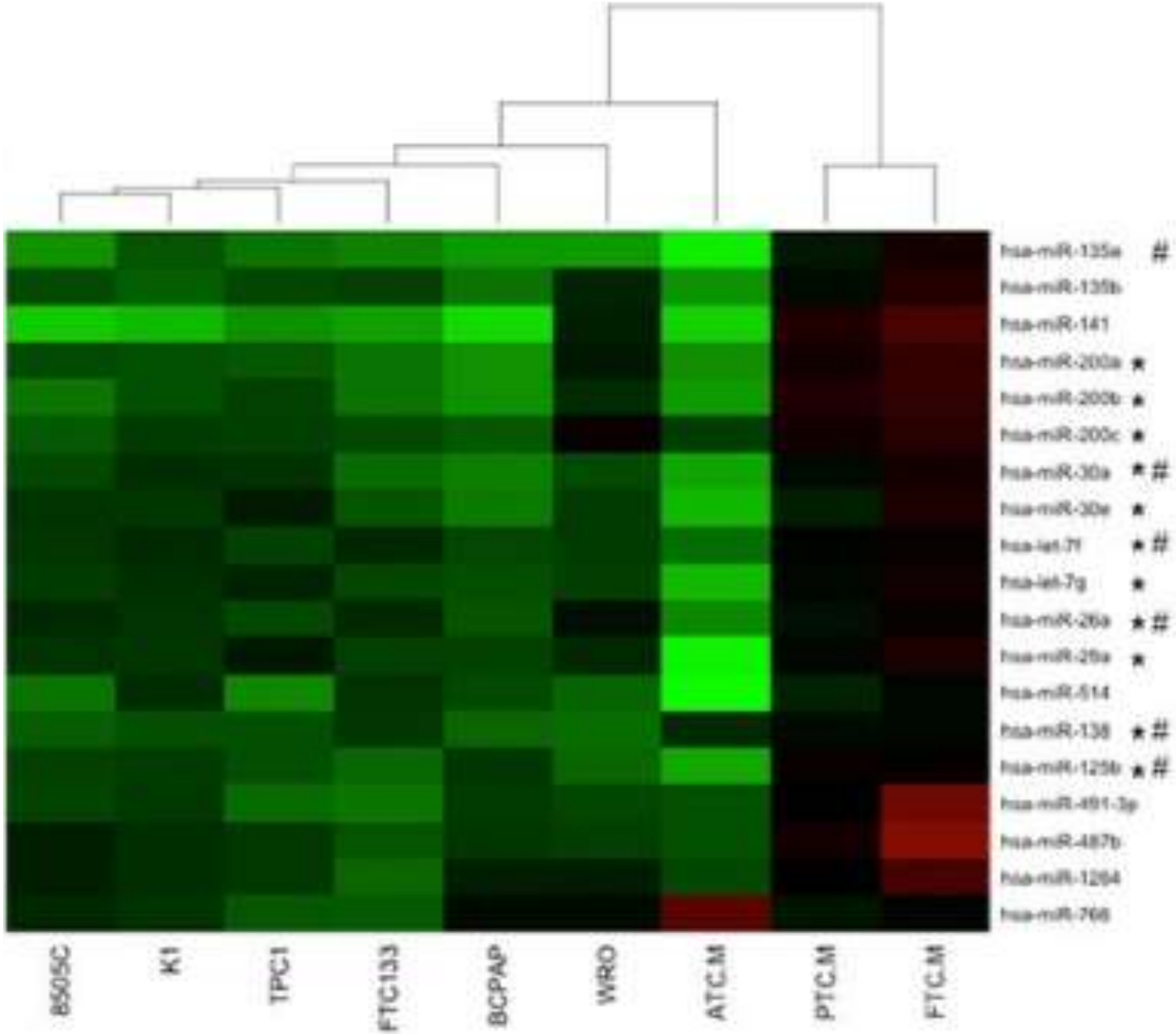
Microarray data: Red and green

Keep in mind Color Conventions:

Male/Female: Blue/Red

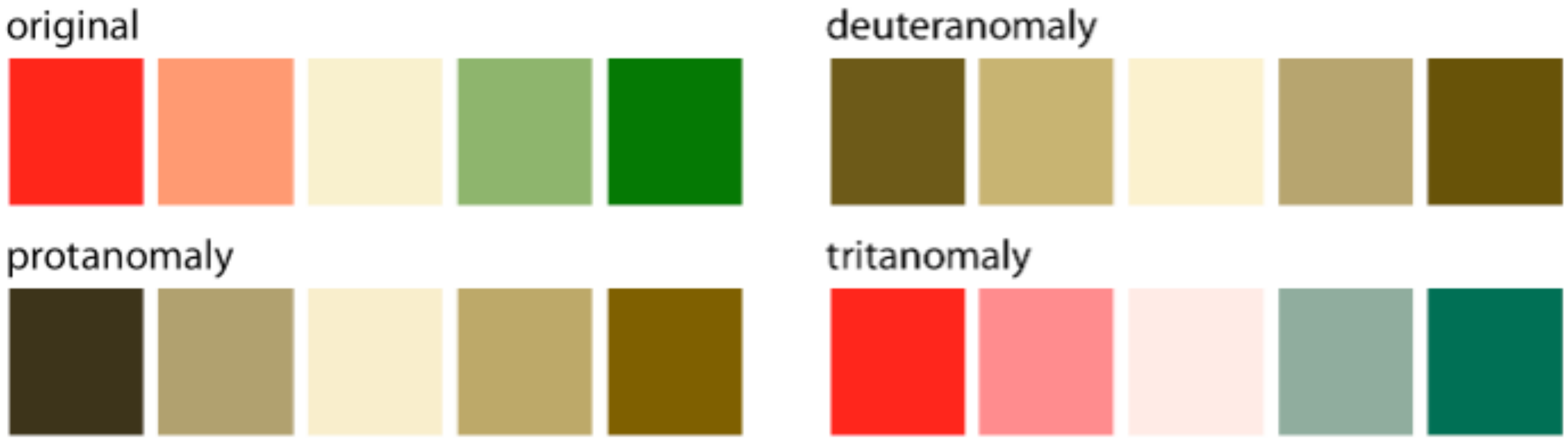
Higher numbers/concentrations: More intense (deeper/more saturated) colors

Color Scheme Example: Heatmap based on Microarray Dataset

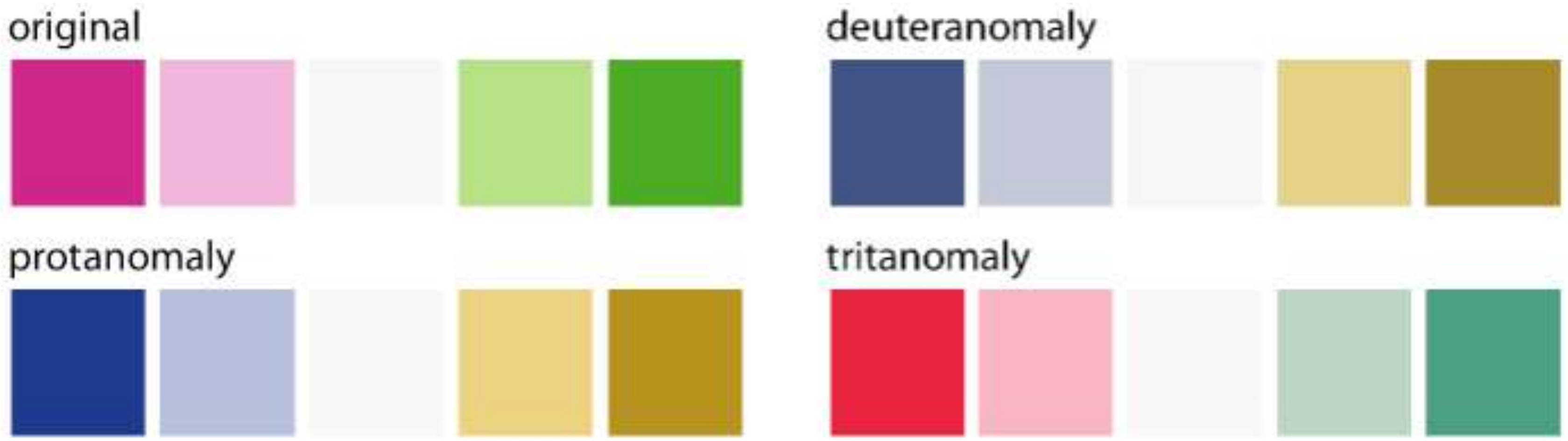


Floor et al, 2014, PloS One

Considering Convention vs Accessibility:



Colorblind-friendly alternative:

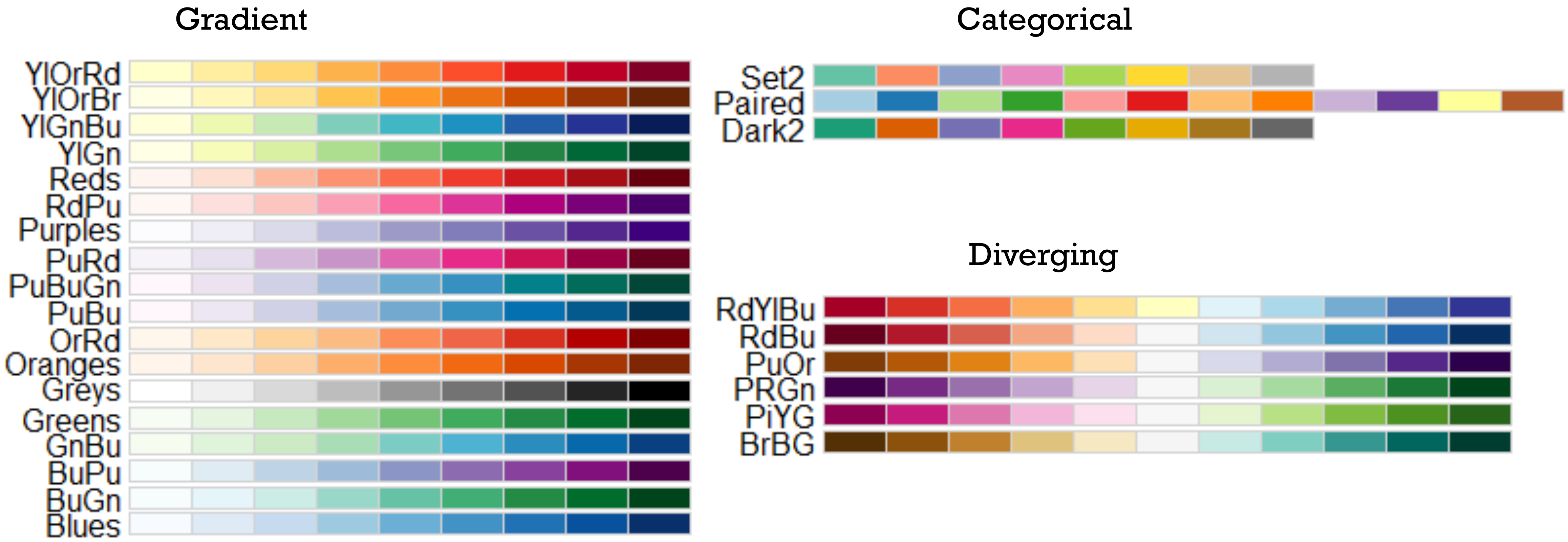


https://serialmentor.com/dataviz/color-pitfalls.html

Colorblind-Friendly Palette Options:

To see a range of different colorblind-friendly palettes already present in the **RColorBrewer** package, type the following into RStudio:

```
library(RColorBrewer)
display.brewer.all(colorblindFriendly = T)
```

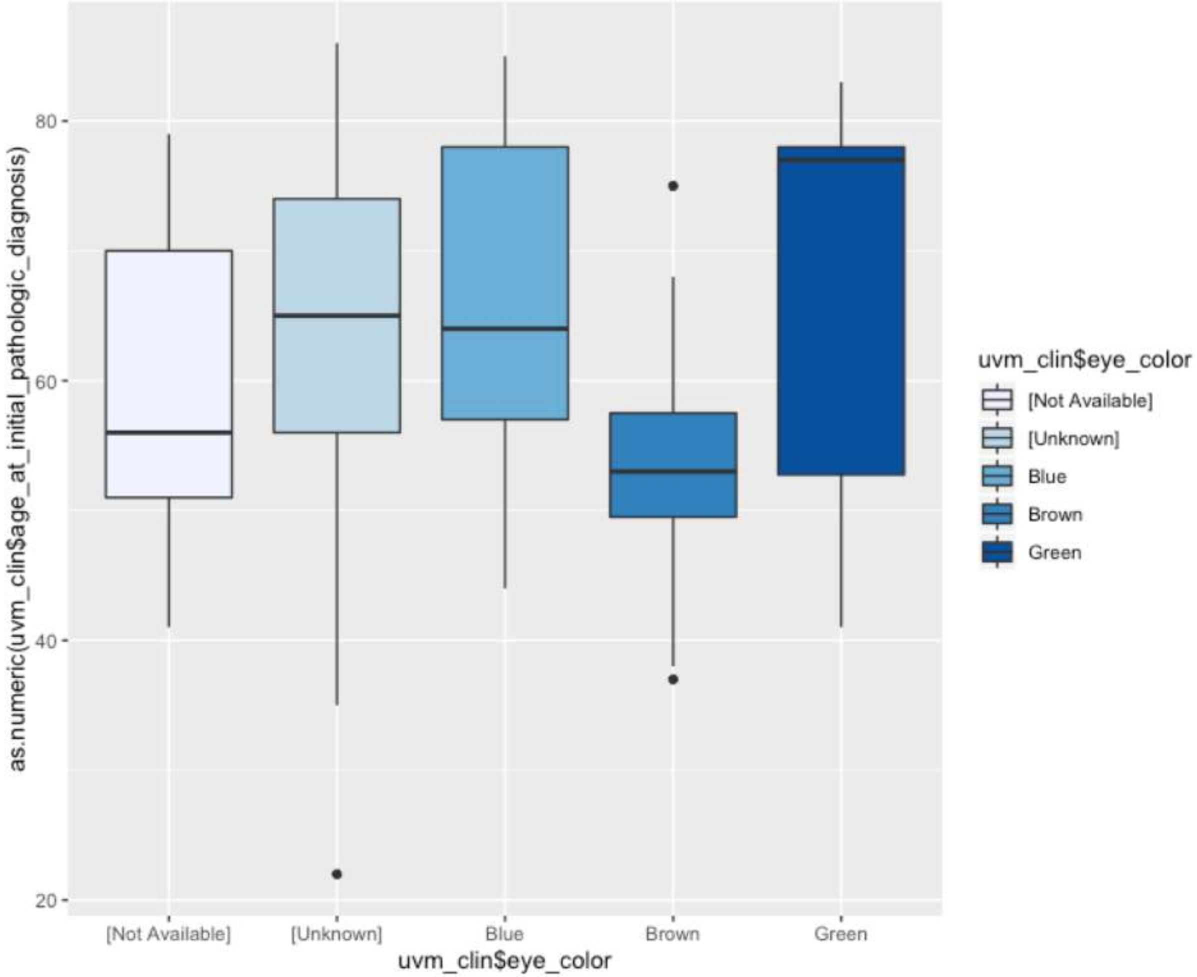


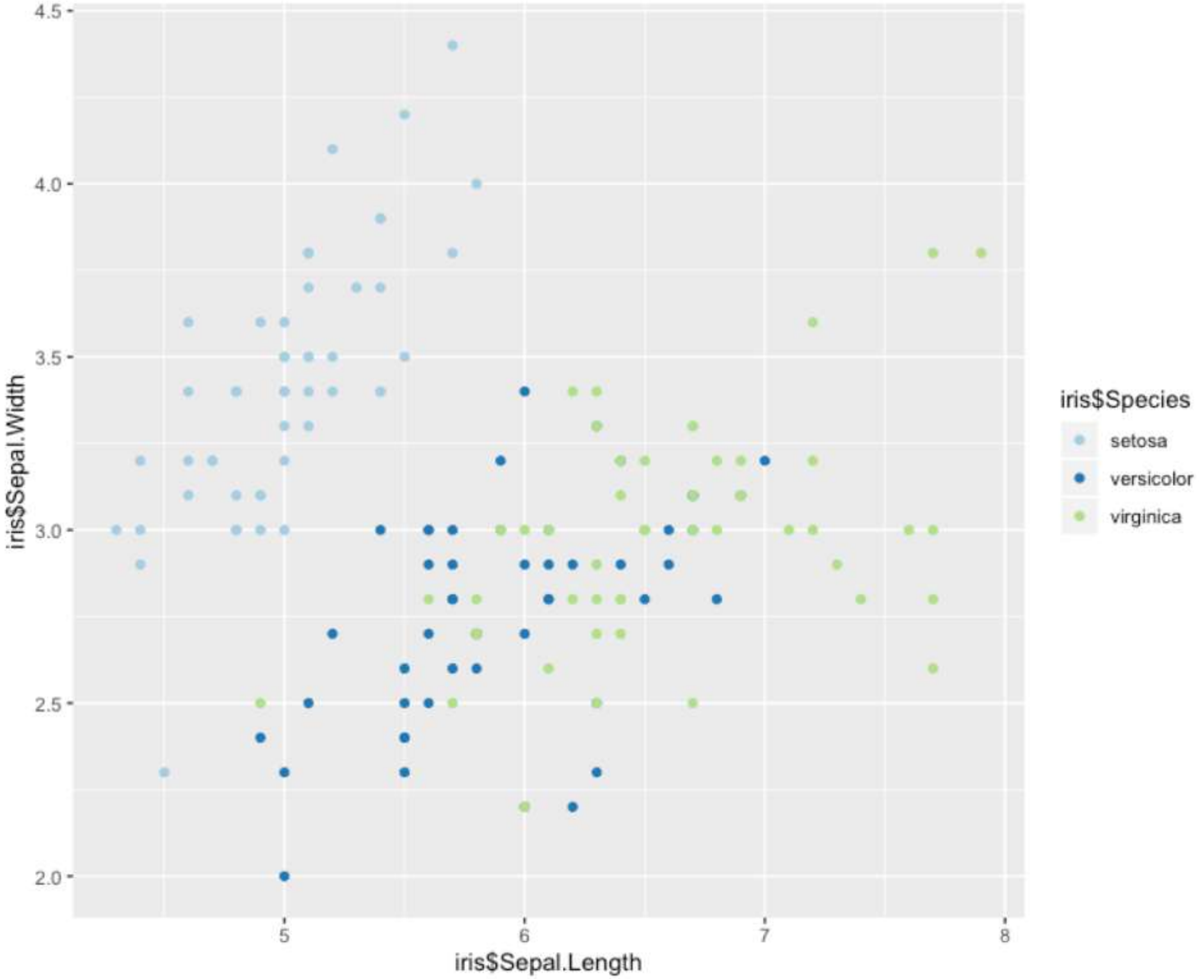
How to apply RColorBrewer package:

1. Edit the aes to add fill or color, as well as what feature you're basing the color scheme on
2. Add the appropriate "brewer" to your ggplot command

- scale_fill_brewer() for box plot, bar plot, violin plot, dot plot, etc
- scale_color_brewer() for lines and points

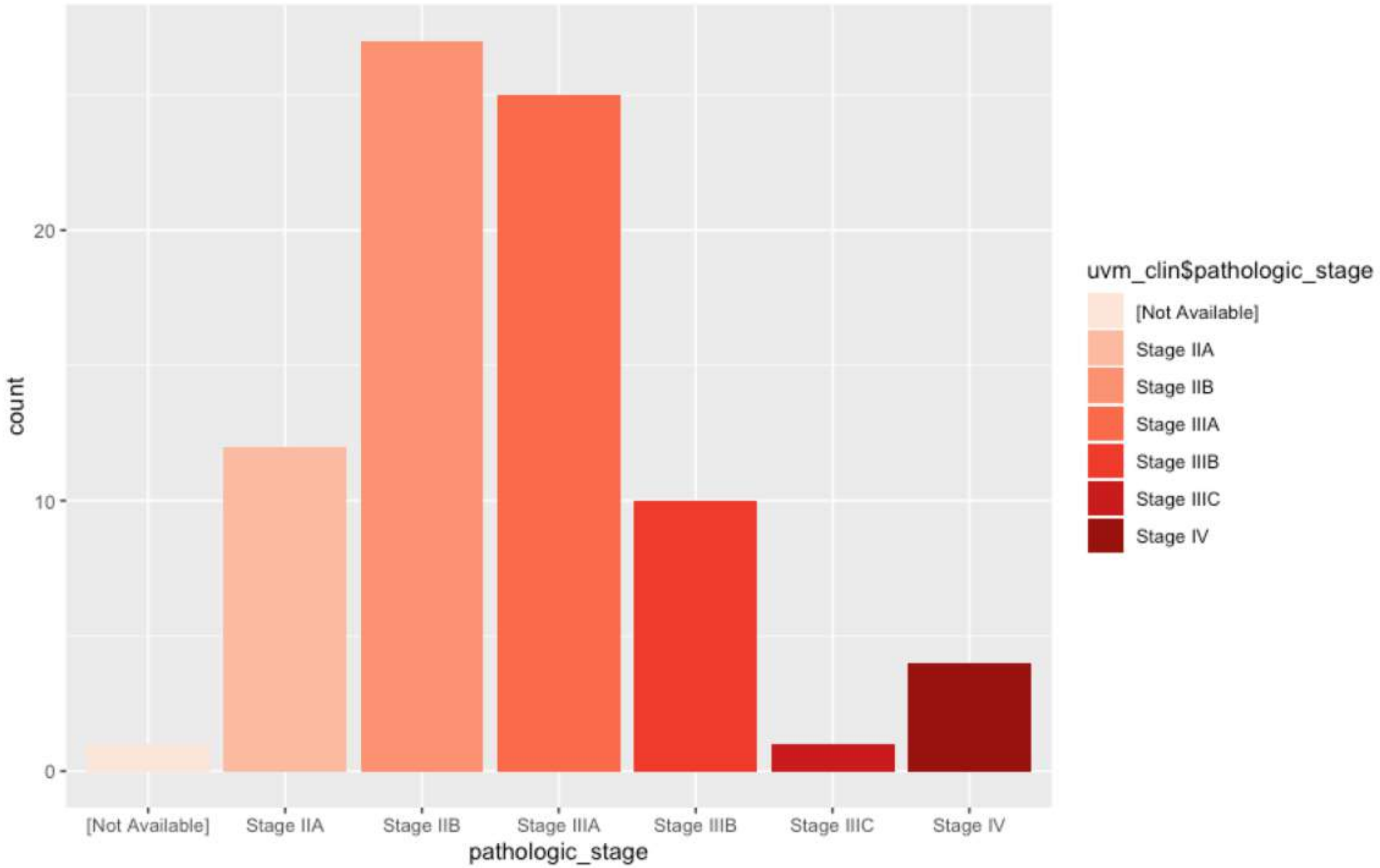
3. Specify your palette of choice: ex: scale_color_brewer(palette="Paired")



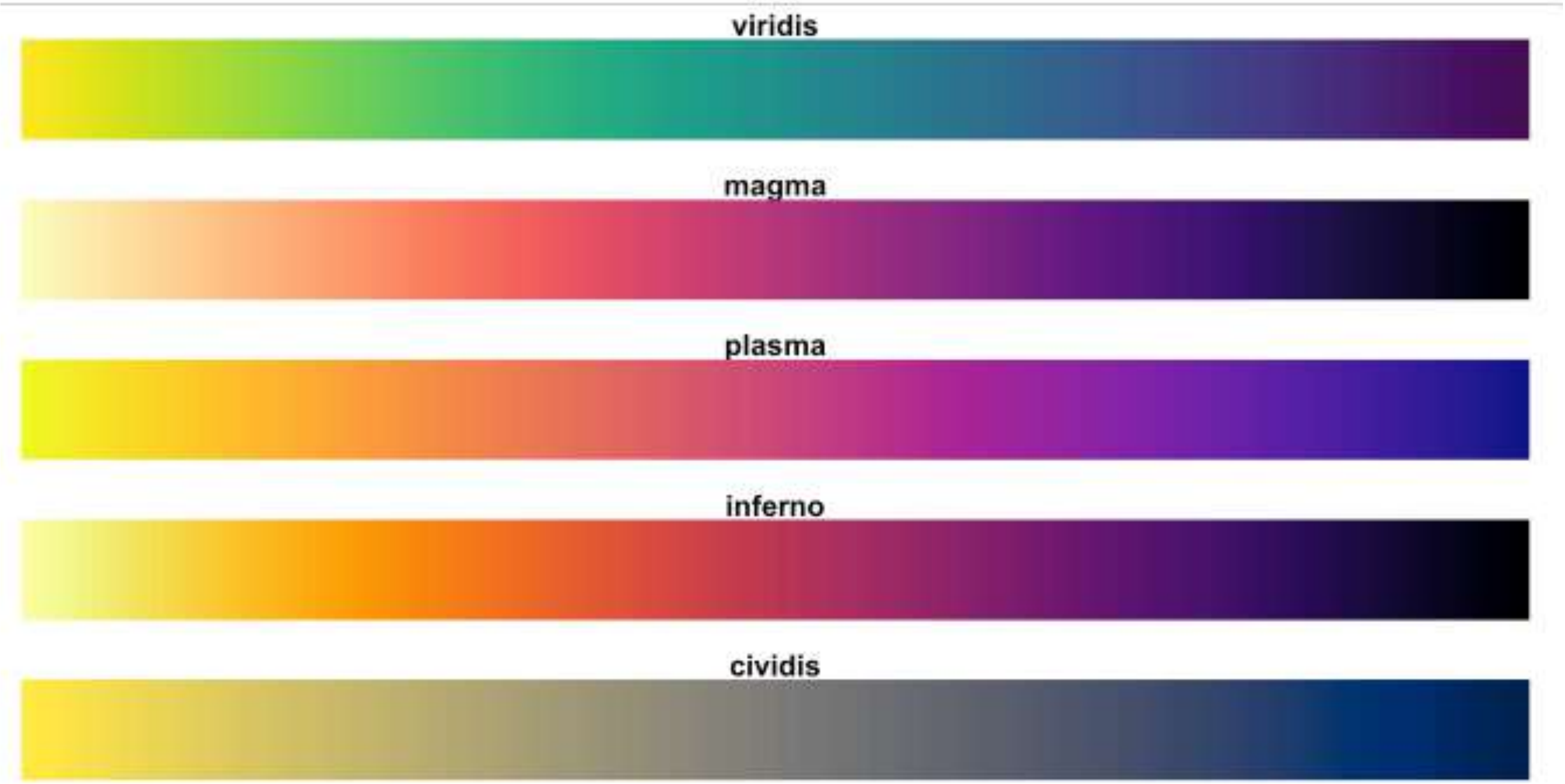


Revisiting gradient colour scheme

- Using Rcolorbrewer!



Colorblind-Friendly Palette Options:



How to apply viridis:

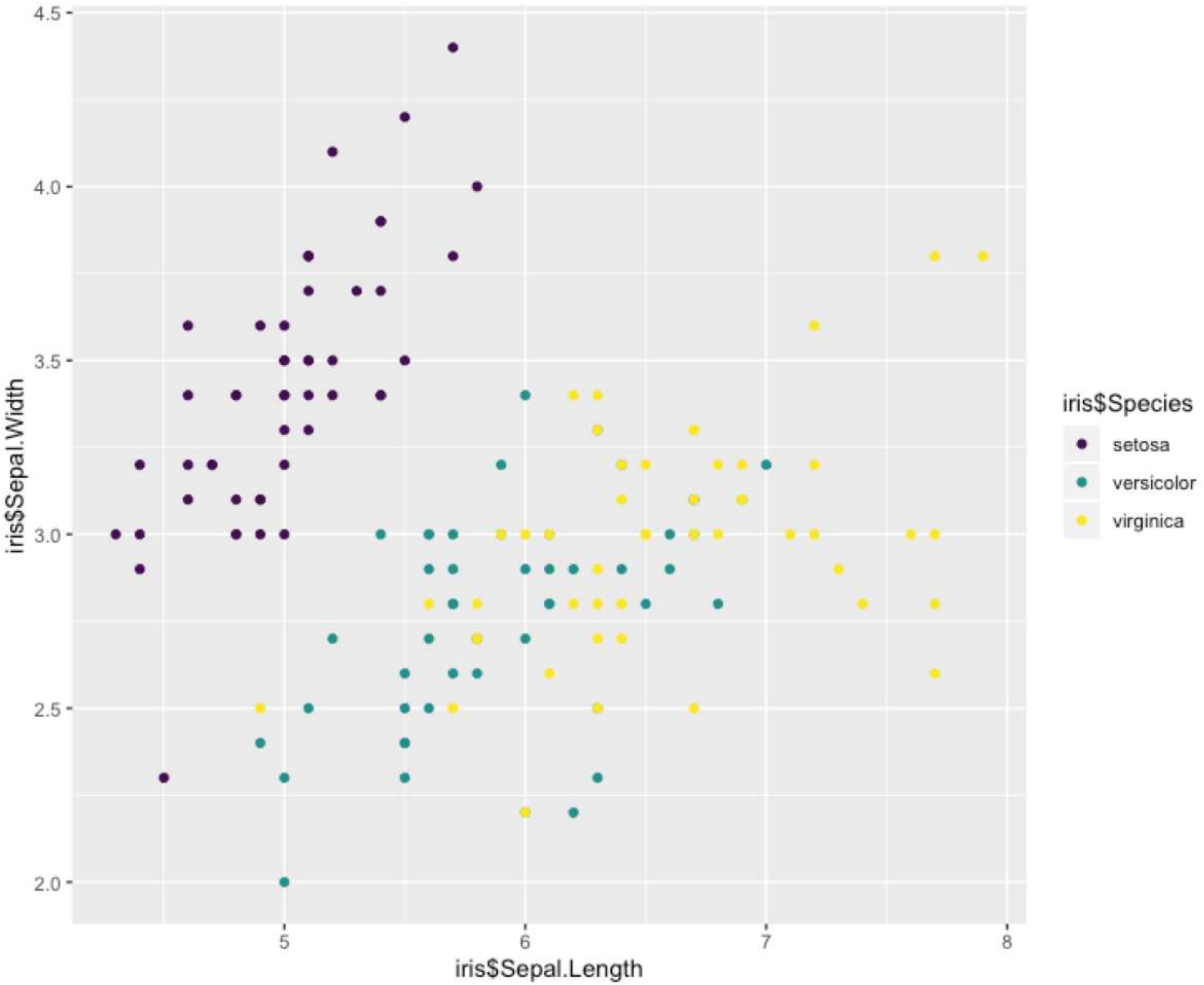
ggplot(iris, aes(iris$Sepal.Length, iris$Sepal.Width, color=iris$Sepal.Length))
+geom_point()+scale_color_viridis(discrete=TRUE)

1. Edit the aes to add fill or color, as well as what feature you're basing the color scheme on
2. Add the appropriate "brewer" to your ggplot command

•scale_fill_viridis() for box plot, bar plot, violin plot, dot plot, etc

•scale_color_viridis() for lines and points

3. If the feature the color scheme is based on is a discrete value, specify it: scale_color_viridis(discrete=TRUE)
4. Specify your palette of choice: ex: scale_color_viridis(option="inferno")



For beautiful color schemes:

The wesanderson R package also has a very nice selection of palettes for a quick, polished look!

```
devtools::install_github("karthik/wesanderson")
library(wesanderson)
```

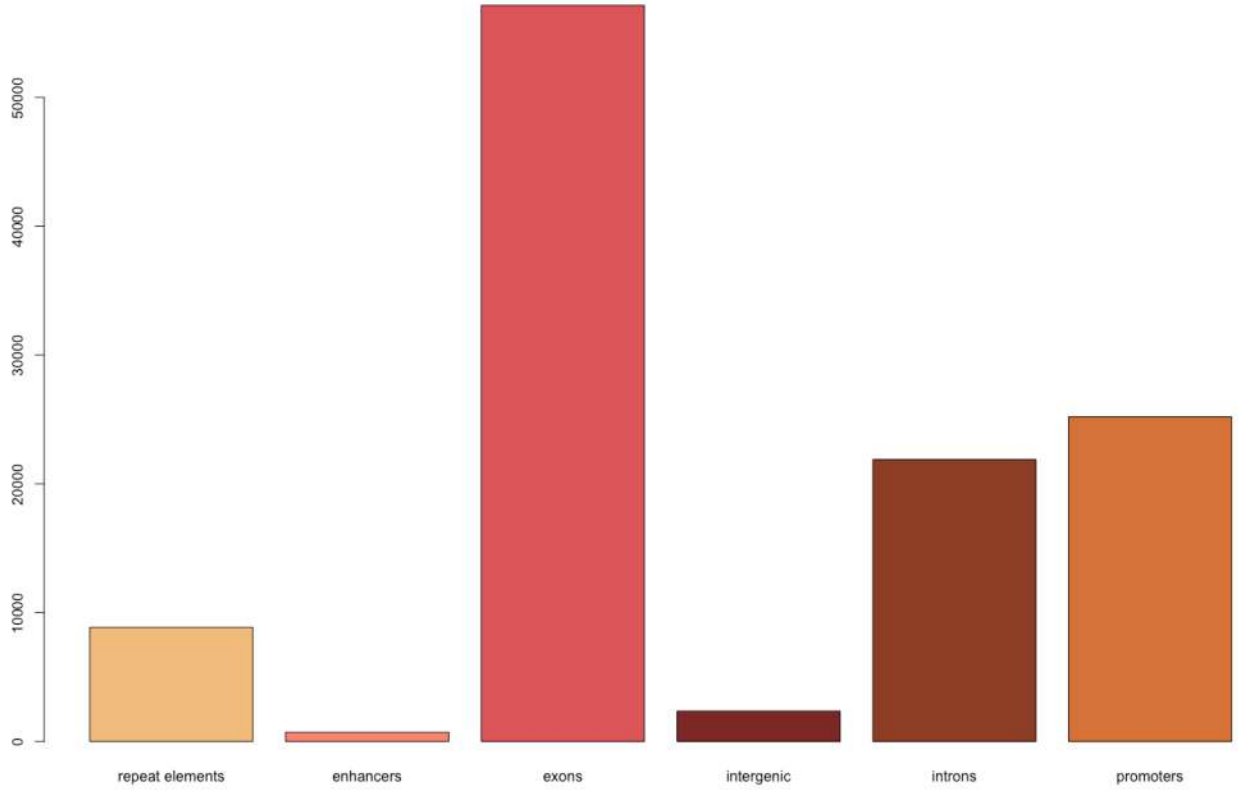
Add to your ggplot command:

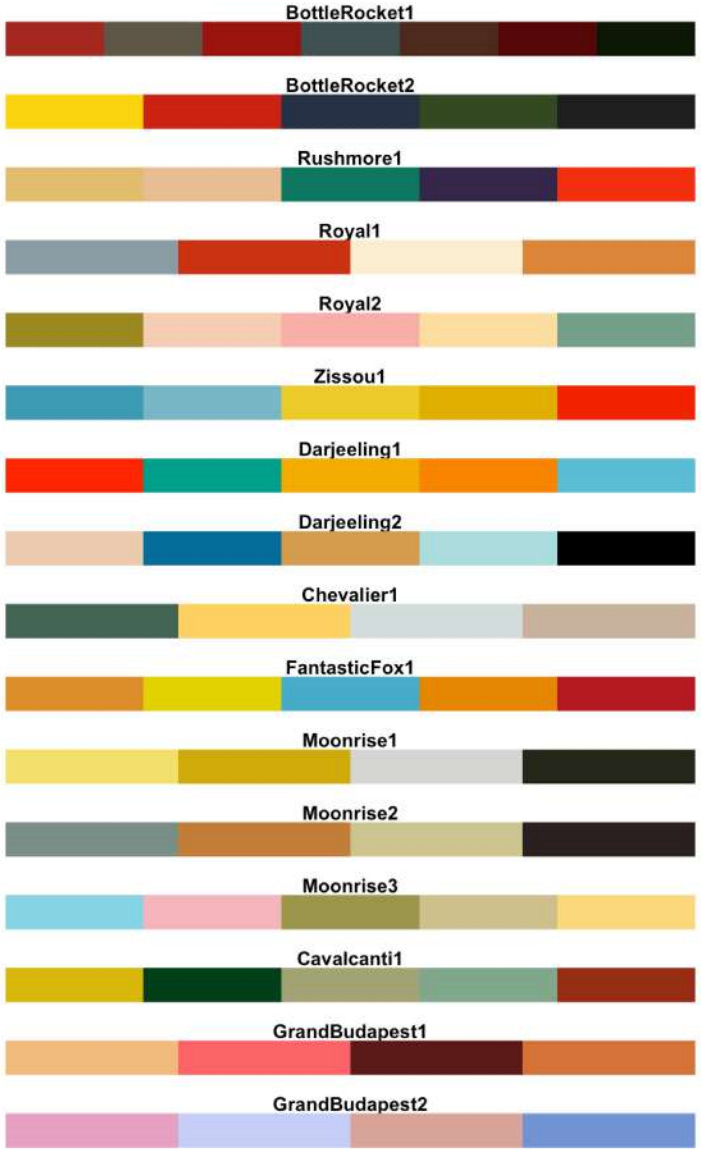
For discrete colours:

```
+ scale_fill_manual(values = wes_palette("GrandBudapest1", n = 3, “discrete”))
```

For gradient colours:

```
+ scale_fill_manual(values = wes_palette("GrandBudapest1", n = 6, “continuous”))
```





Exercise 3:

Using the data provided in either uvm_counts or uvm_clin (or both!), generate a plot that requires fill for colour and one that requires color for colour to answer a biological question in the data that interests you!

Try both 1) specifying the colours independently from the color value table

And

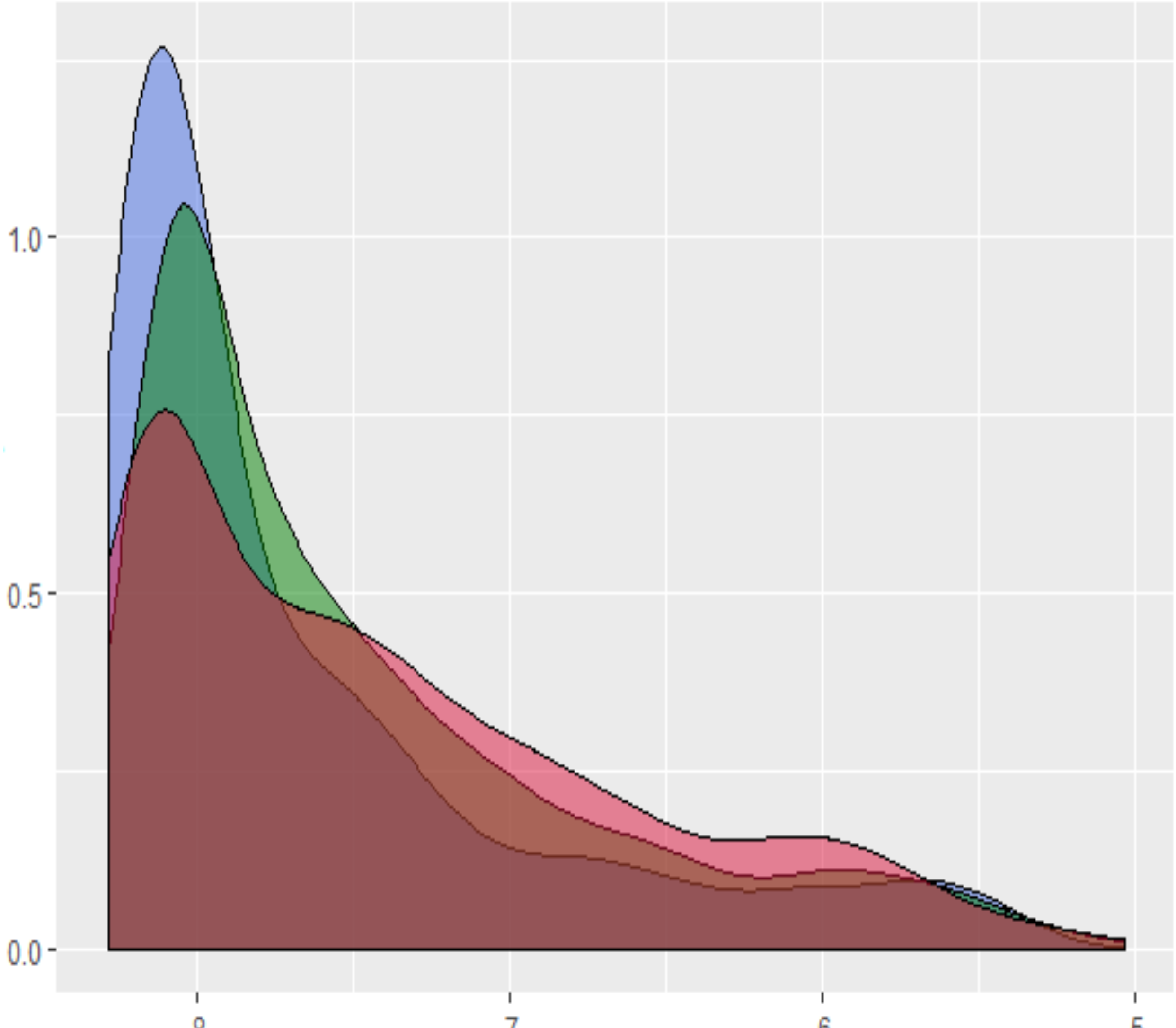
2) Using the pre-built palettes

Keep in mind the principles of color theory that we discussed earlier!

What are the pros and cons of either color scheme strategy?

Transparency:

Add to your ggplot command where you say the type of plot you want to make
alpha= , followed by a number between 0 (completely transparent) and 1 (completely opaque).
Ex:
+ geom_density(alpha=0.4)



Exercise 4:

Compare the distribution of ages at which the pathological diagnosis occurred between male and female patients:

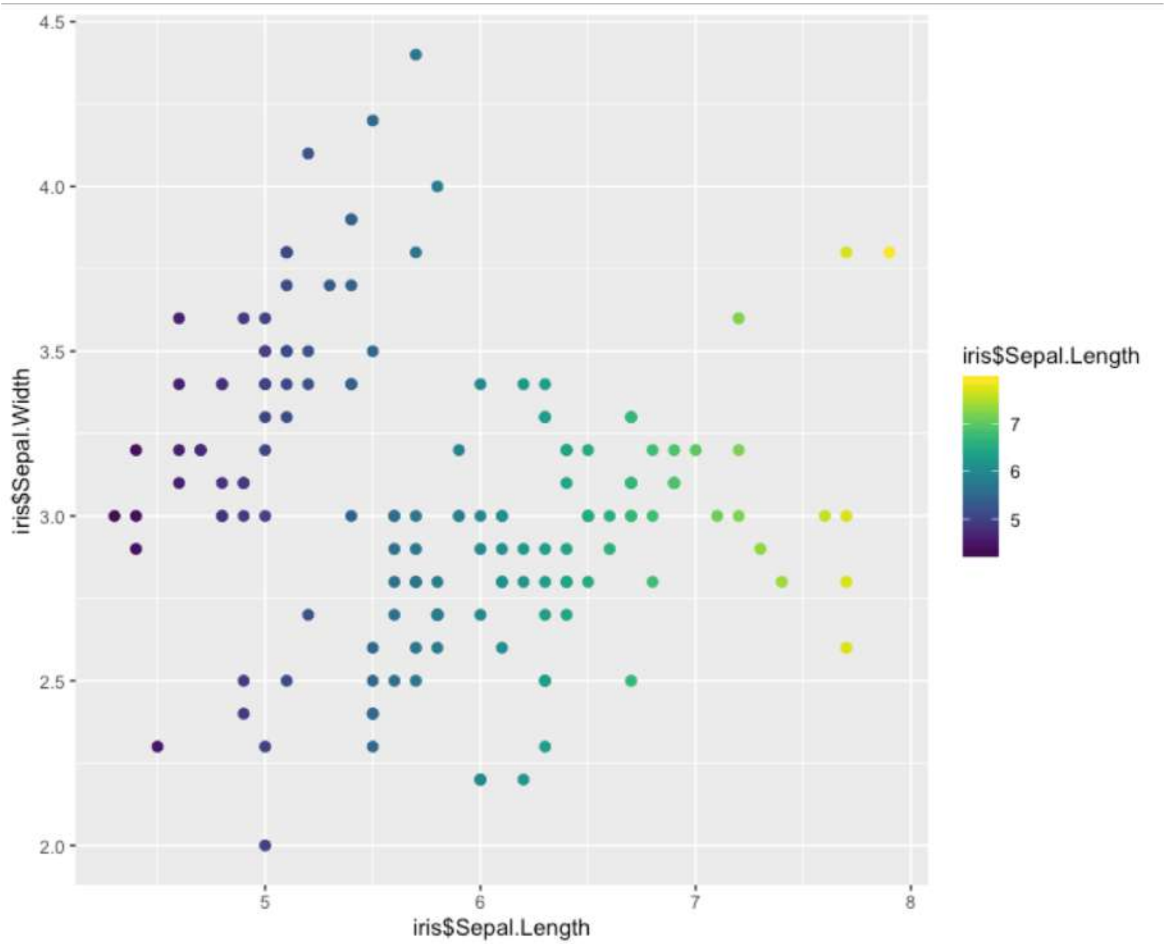
A) Make the appropriate density plot
B) Make use of the transparency and color tools ggplot2 has to improve your plot!

How to change size of points

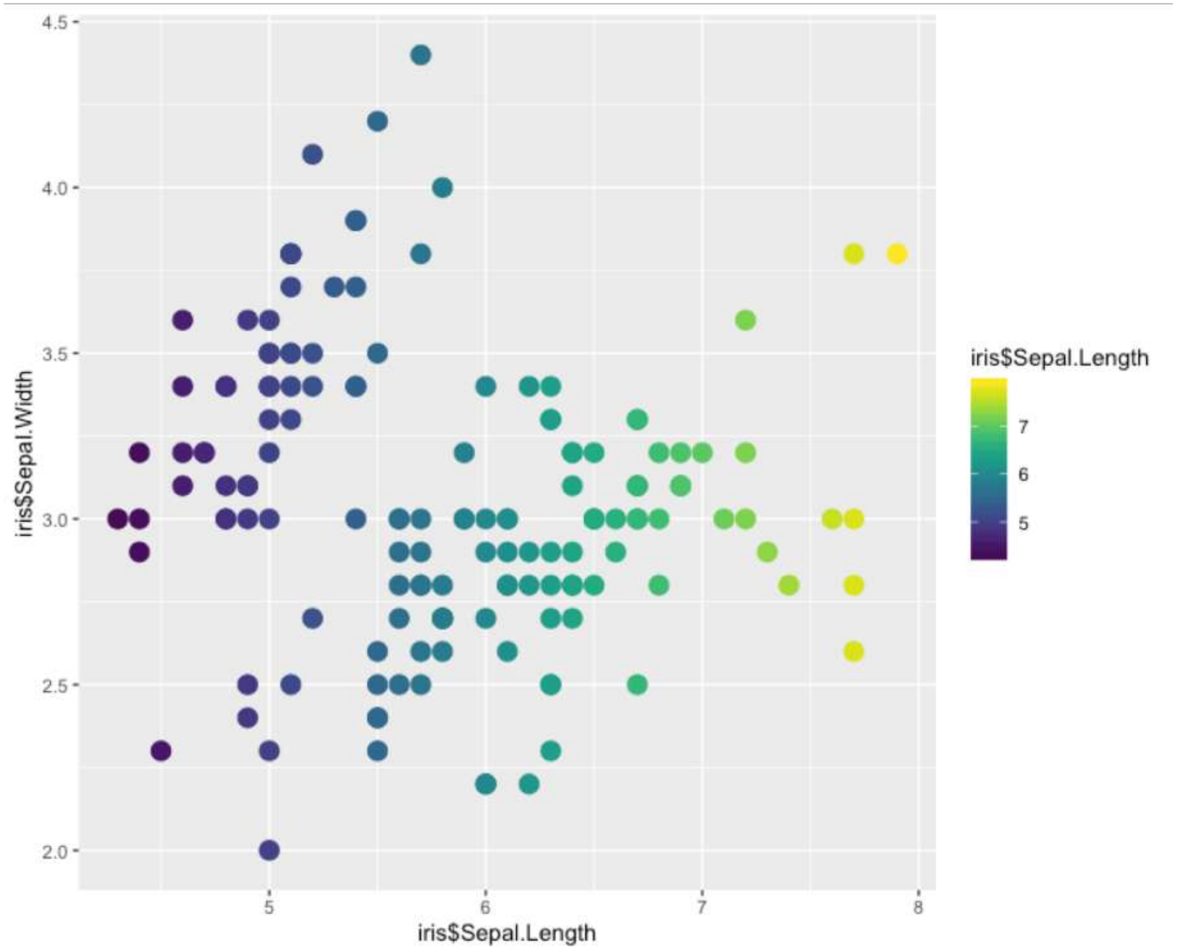
geom_point(size=a)

ggplot(tcga_express, aes(dna_meth, gene_A_exprs)) + geom_point( )

geom_point(size=2)

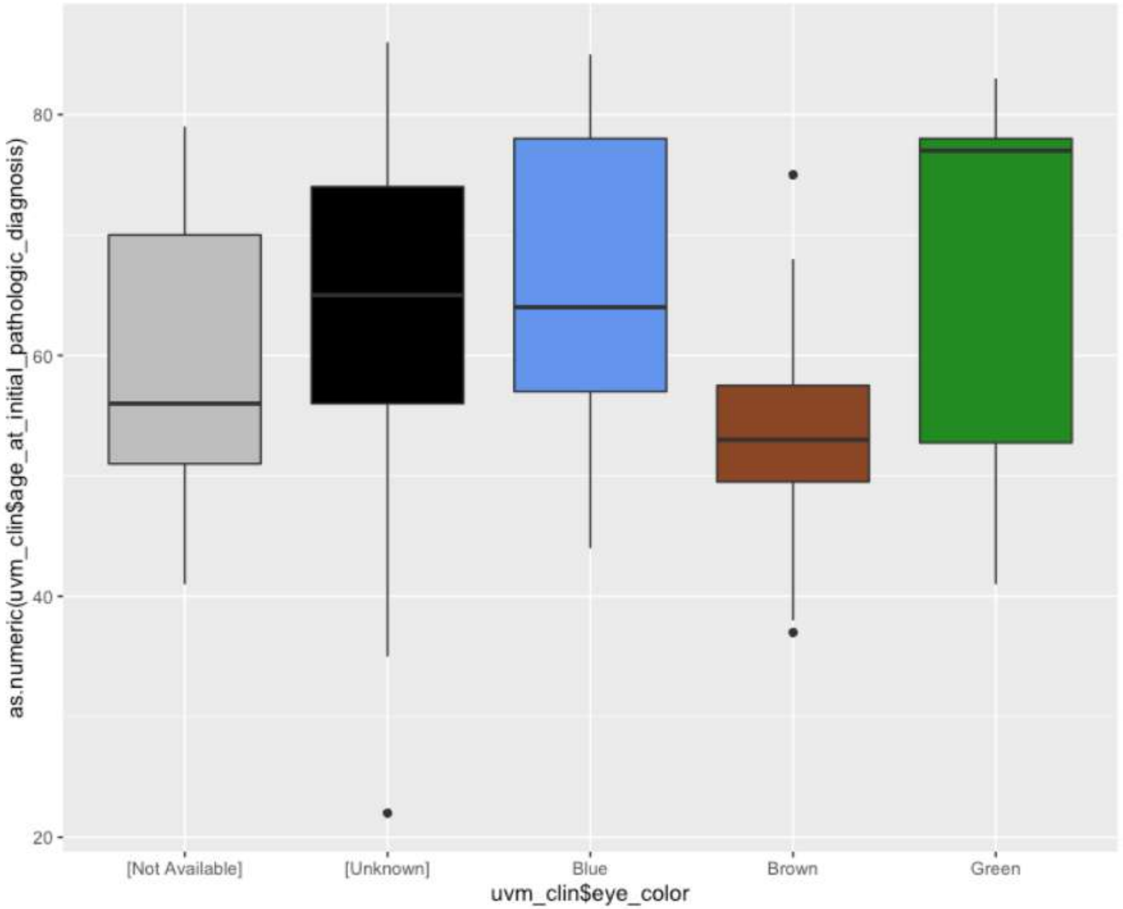


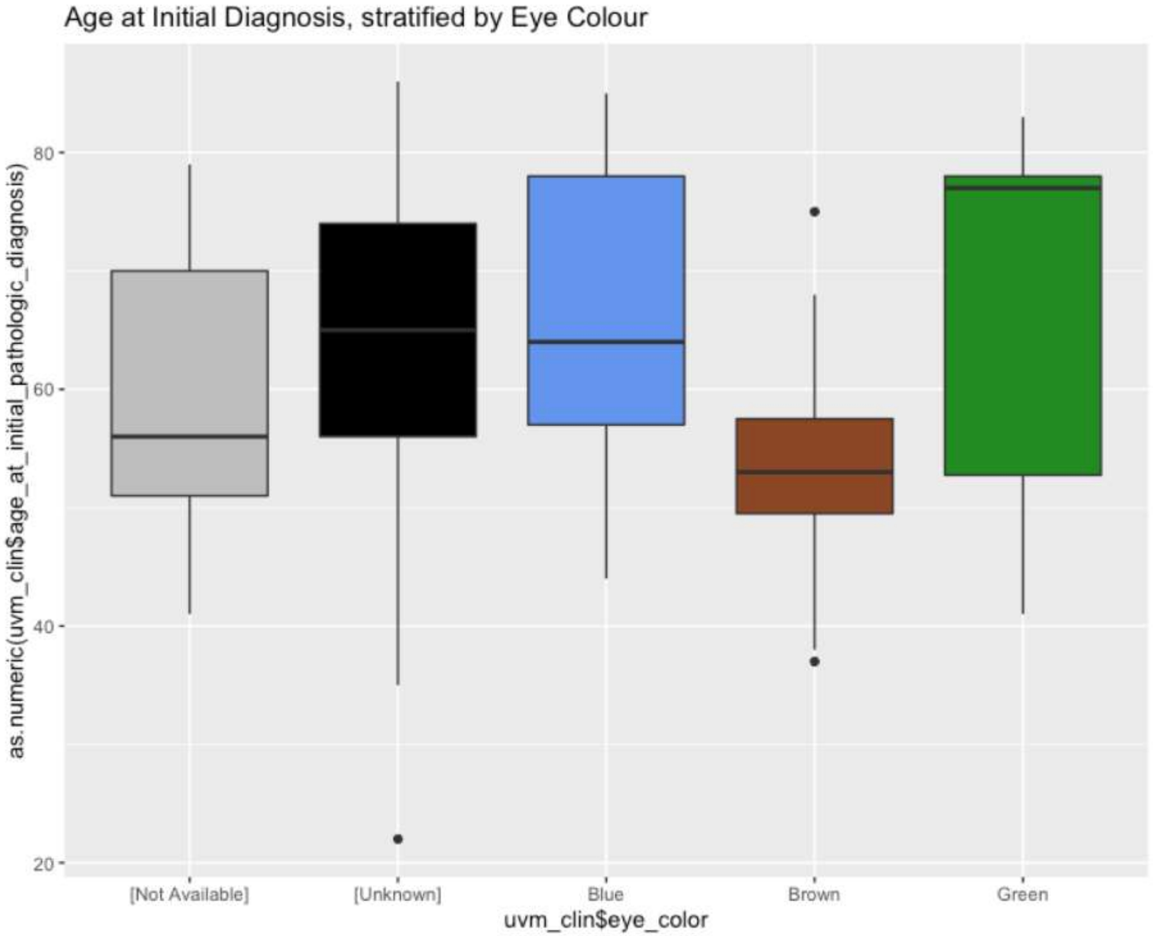
geom_point(size=4)



Adding a title

```
+ggtitle("Age at Initial Diagnosis, stratified by Eye Colour")
```



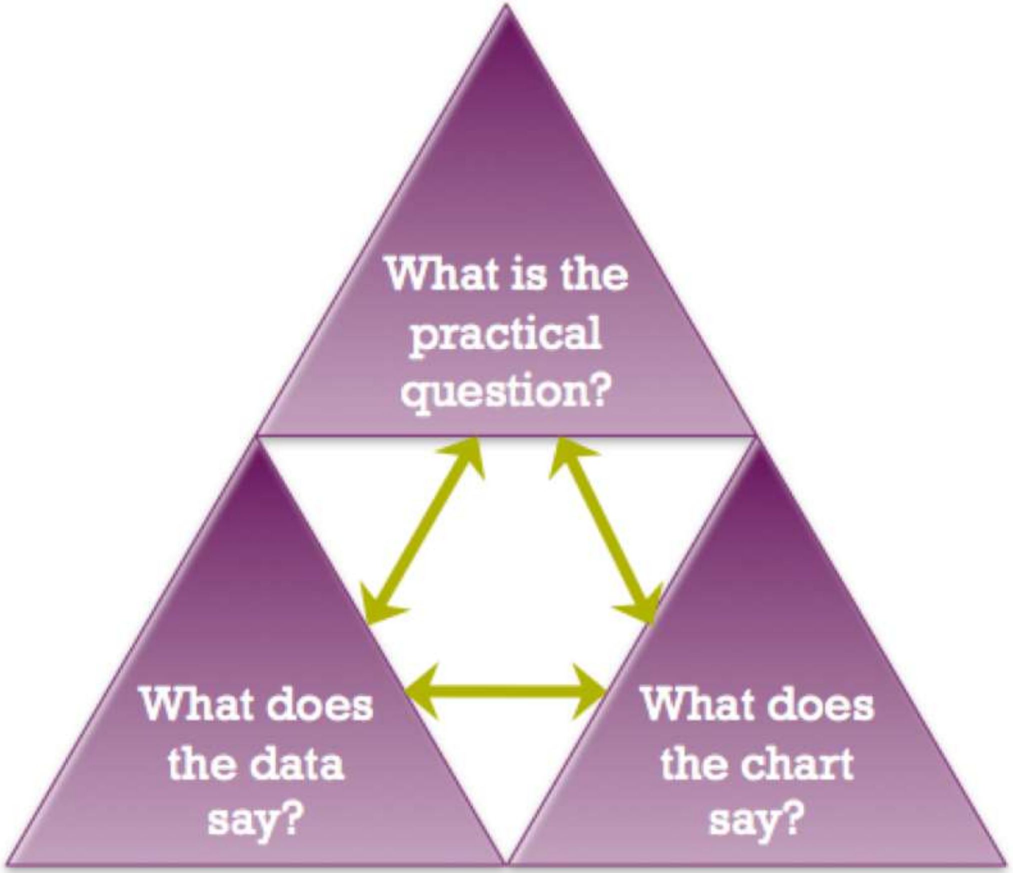


Presentation

Principles of Good Design:

Why do we make plots and figures?
What makes “good” or “bad” design?

Guiding questions in figure design:

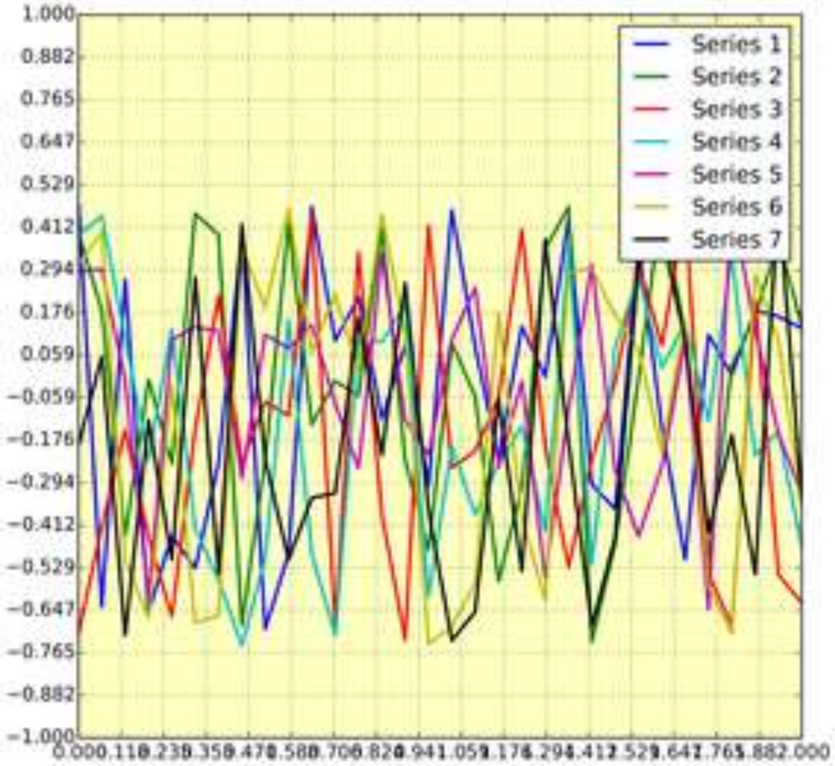


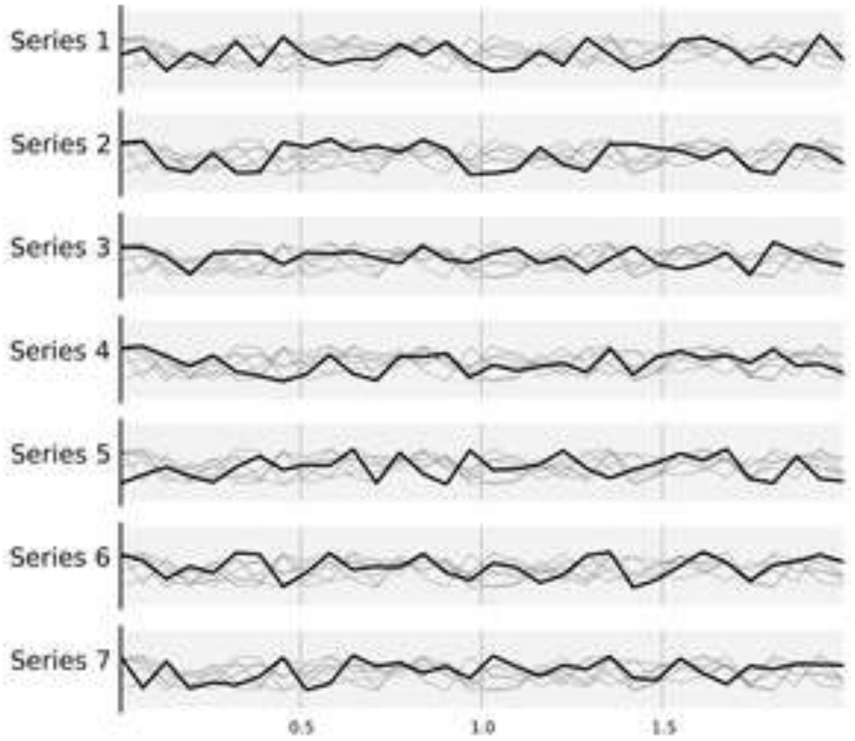
Kaiser Fung, https://junkcharts.typepad.com/junk_charts/junk-charts-trifecta-checkup-the-definitive-guide.html

Exercise 5:

Using the different datasets provided and the functions that we learned today, generate a plot (or plots!) exploring a biological question that you found interesting within the data you have available.

Avoid chartjunk

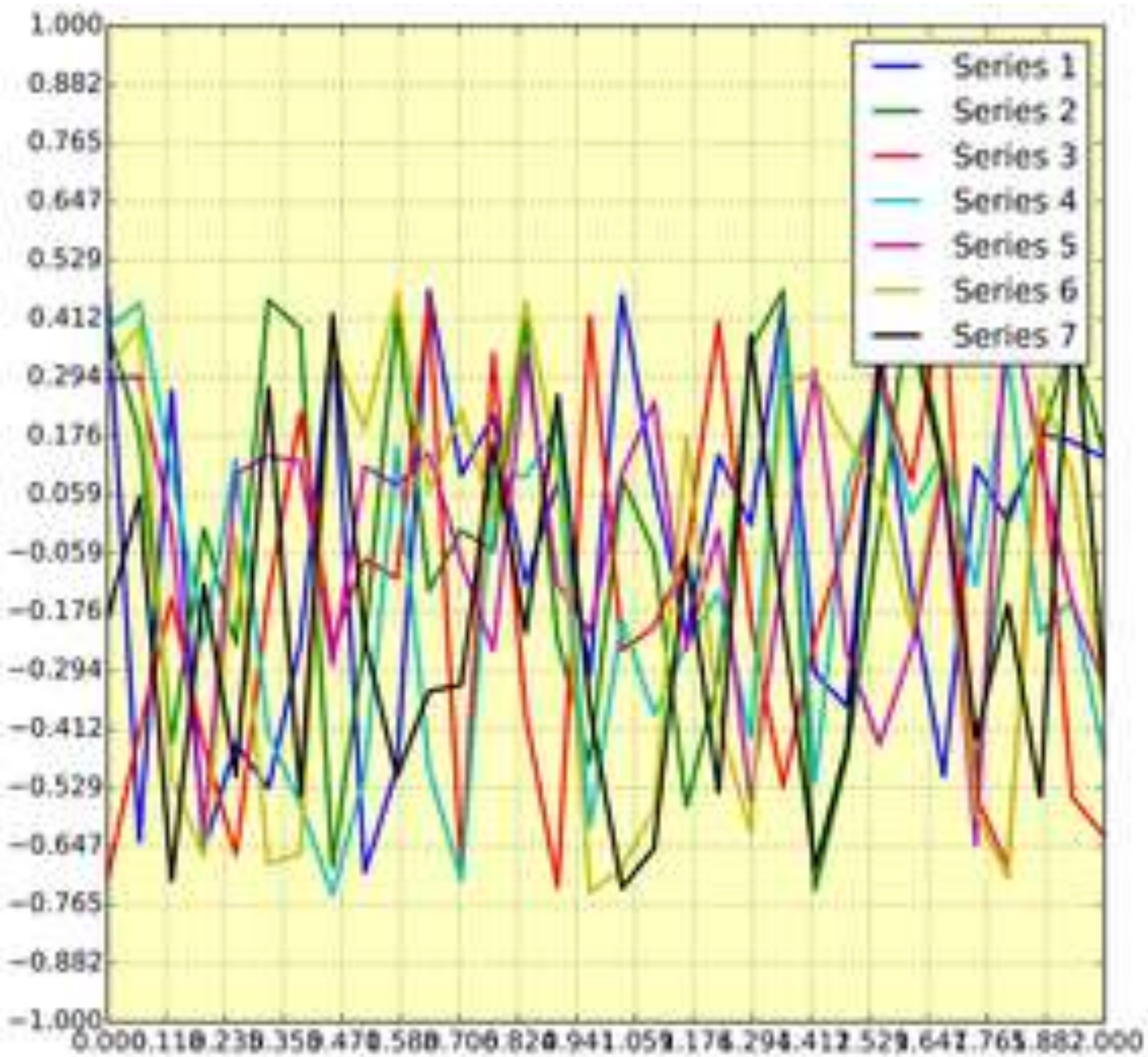




Rougier NP, Droettboom M, Bourne PE (2014) Ten Simple Rules for Better Figures. PLOS Computational Biology 10(9): e1003833. https://doi.org/10.1371/journal.pcbi.1003833
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833

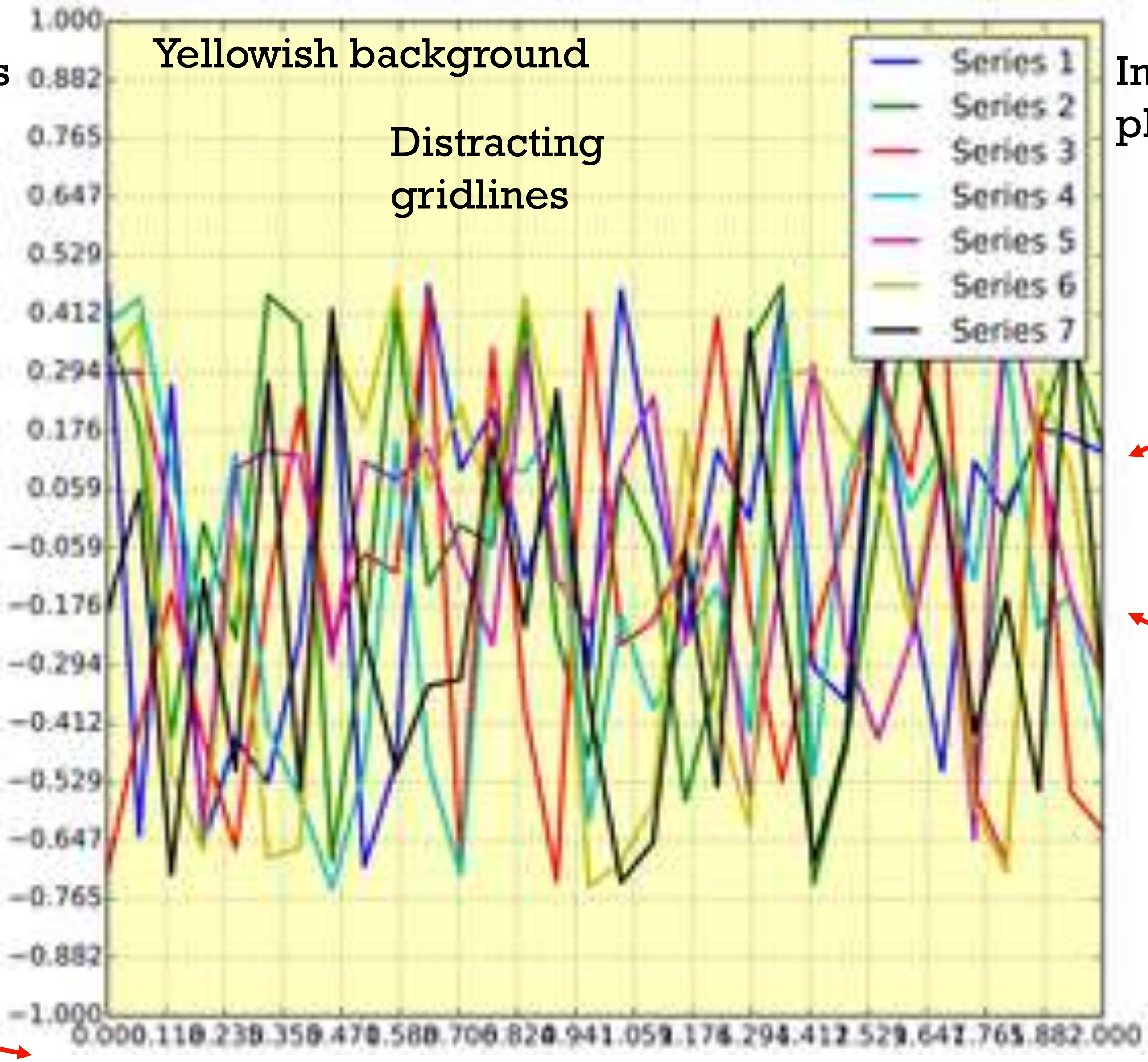
PLOS | COMPUTATIONAL BIOLOGY

Q: What is wrong with this plot?



A: What isn't?

Unhelpful axis increments



Yellowish background

Distracting gridlines

Inconveniently placed legend

Way too many overlapping lines, very dense

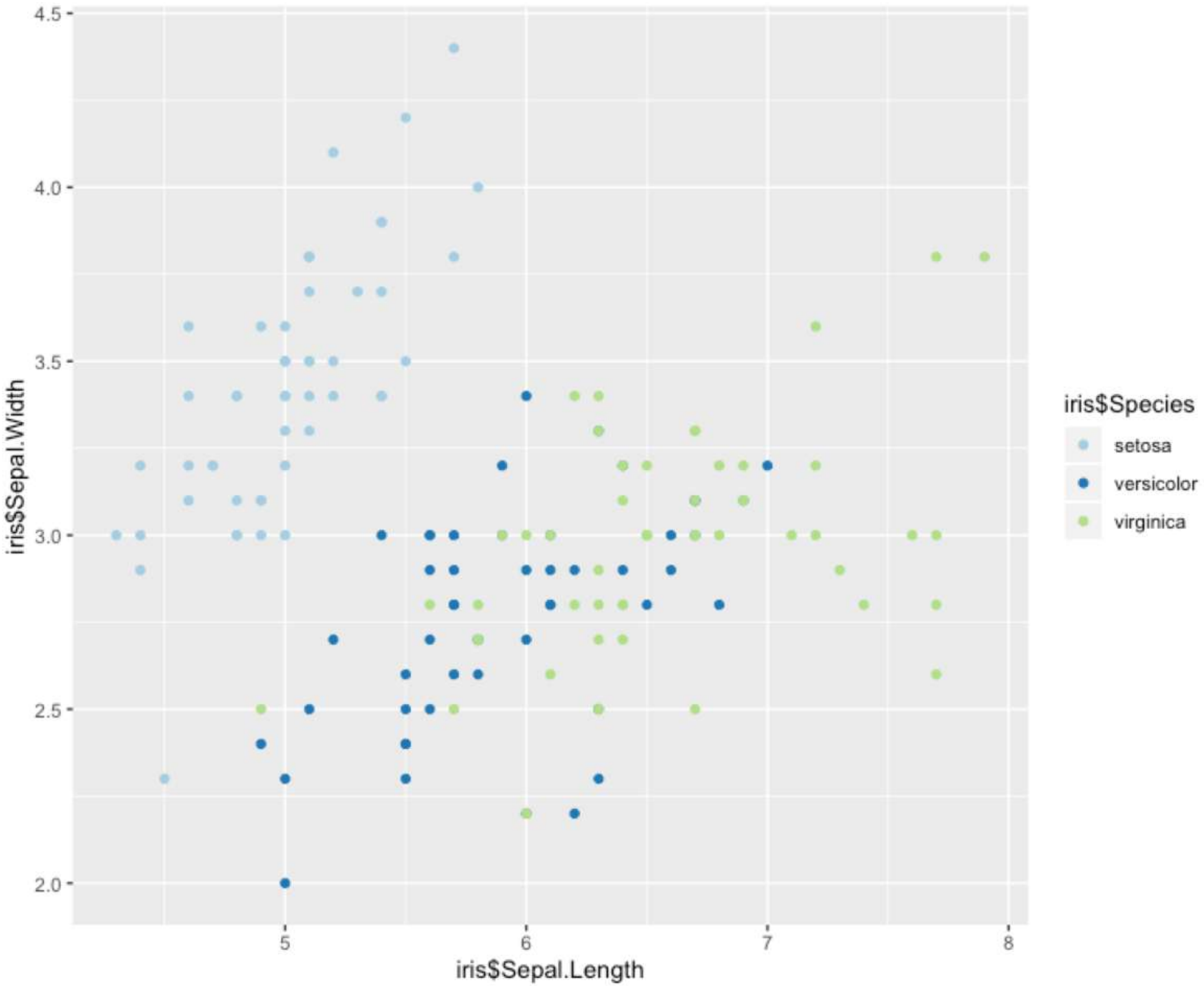
The message is lost due to poor data viz

Too many colors

Axes don't have titles

Axis labels are too crowded

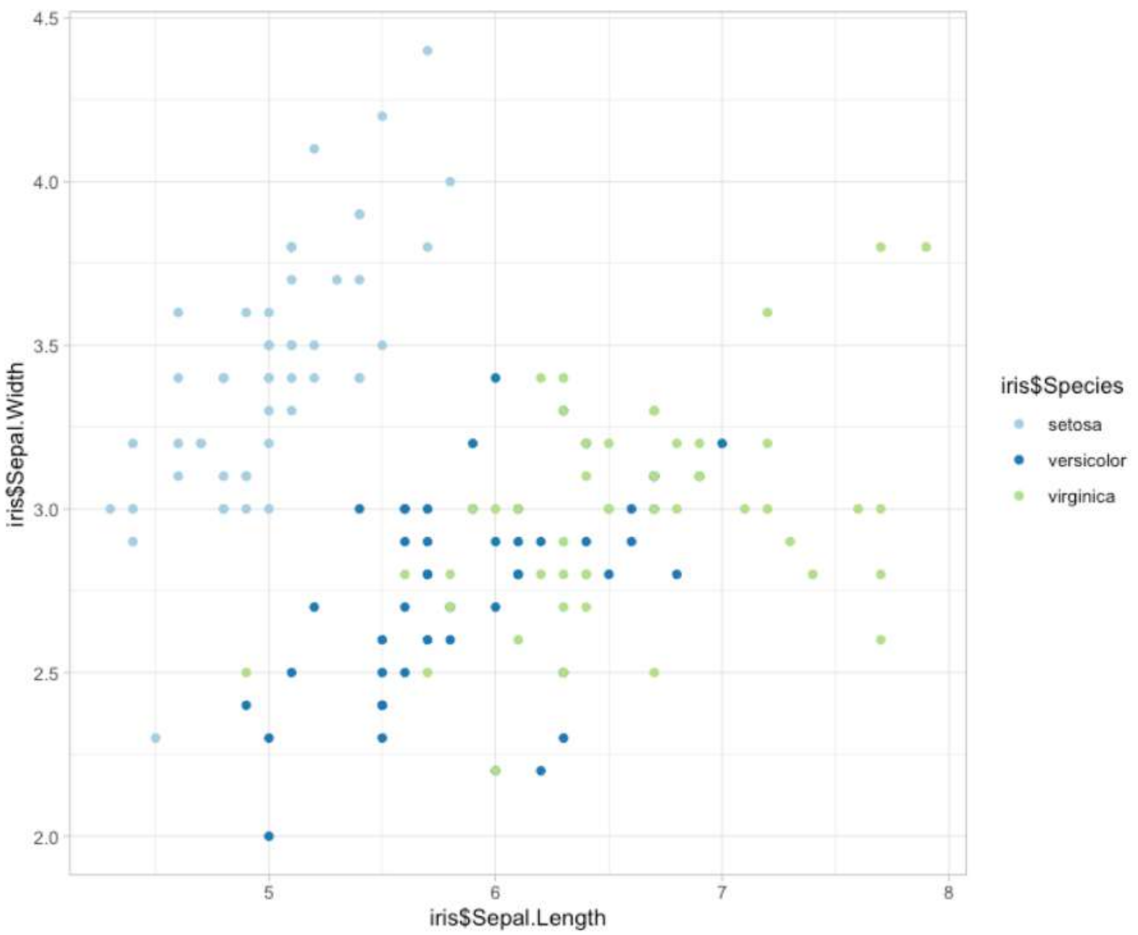
What would you improve with this plot?



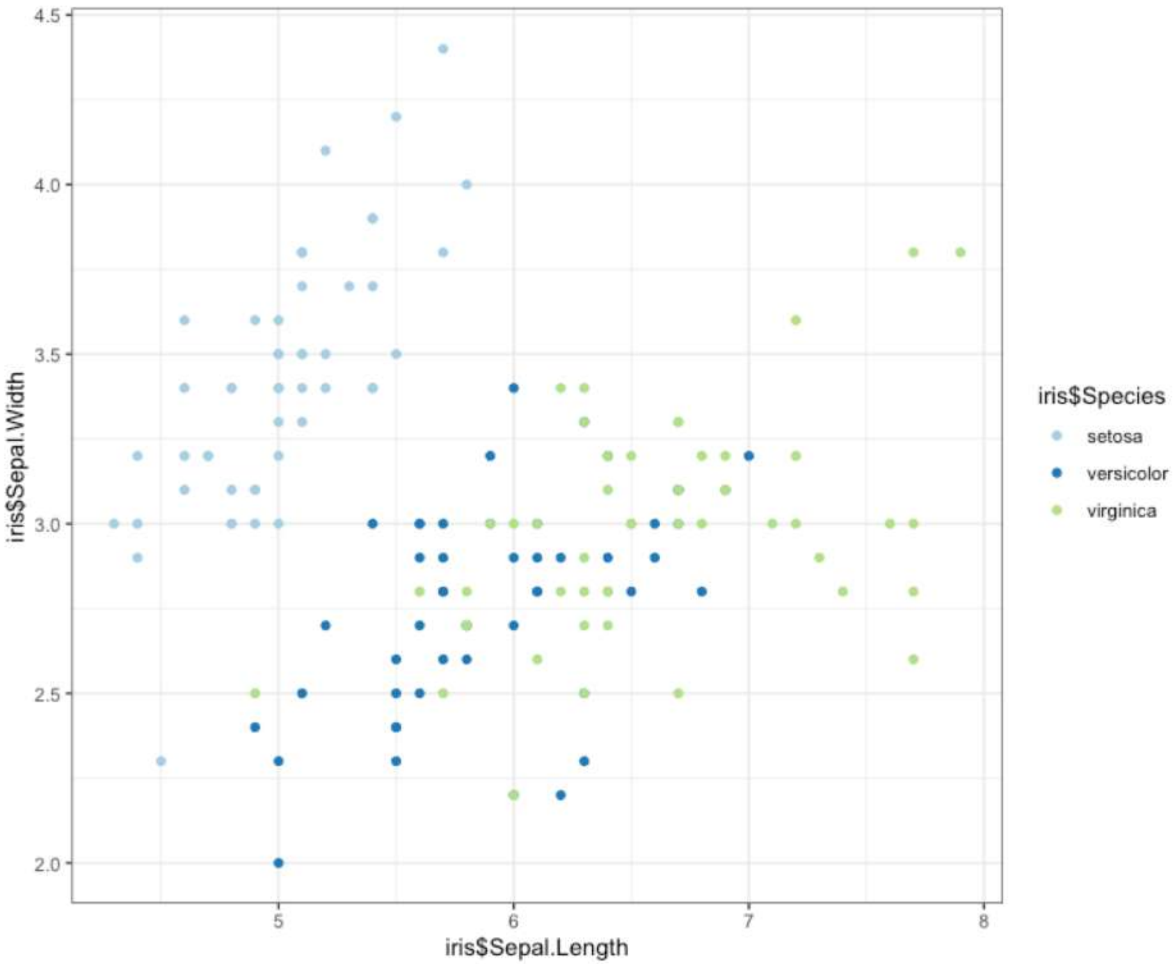
Cleaning up your plot

Adjusting the background

+theme_light()

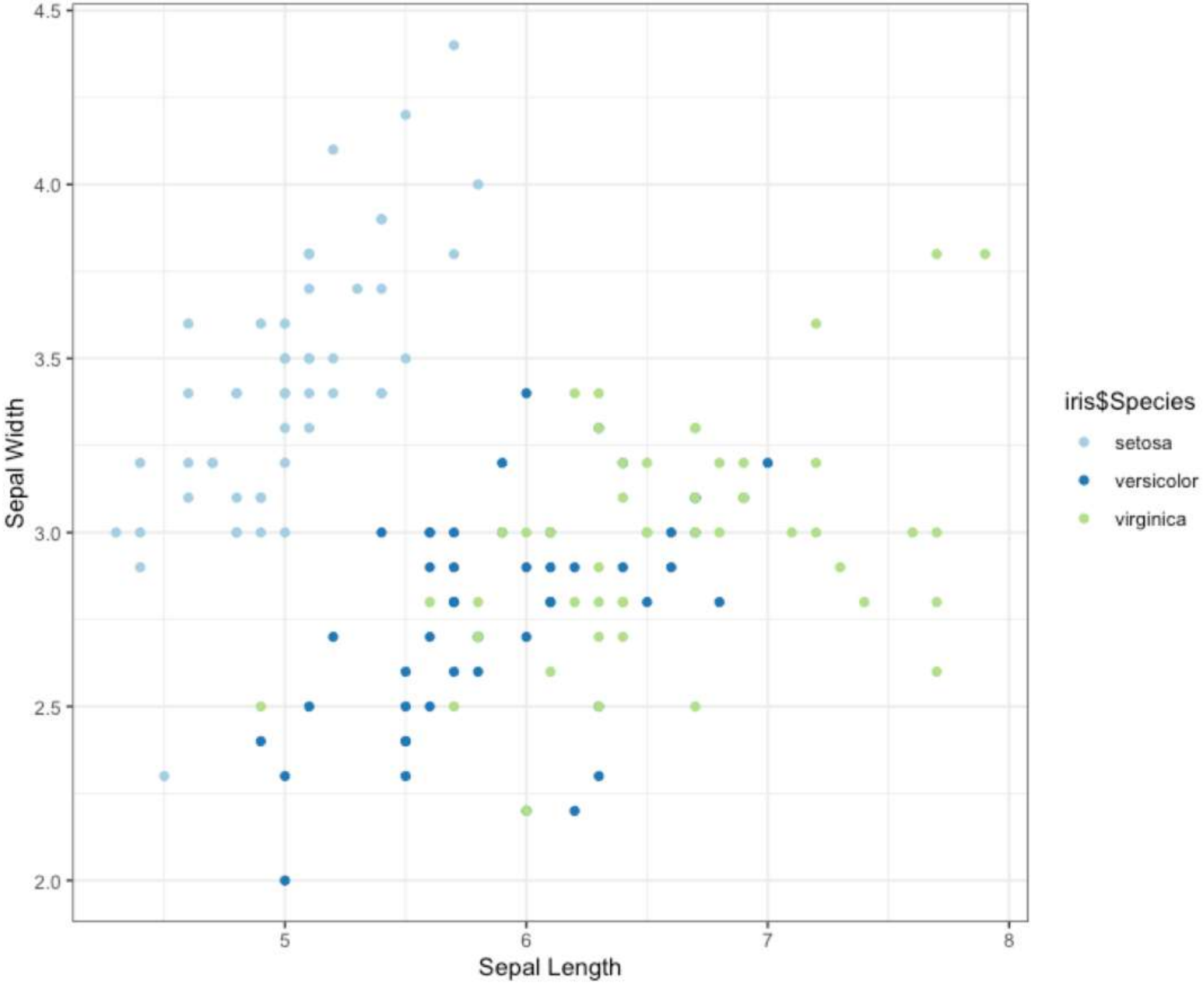


+theme_bw()



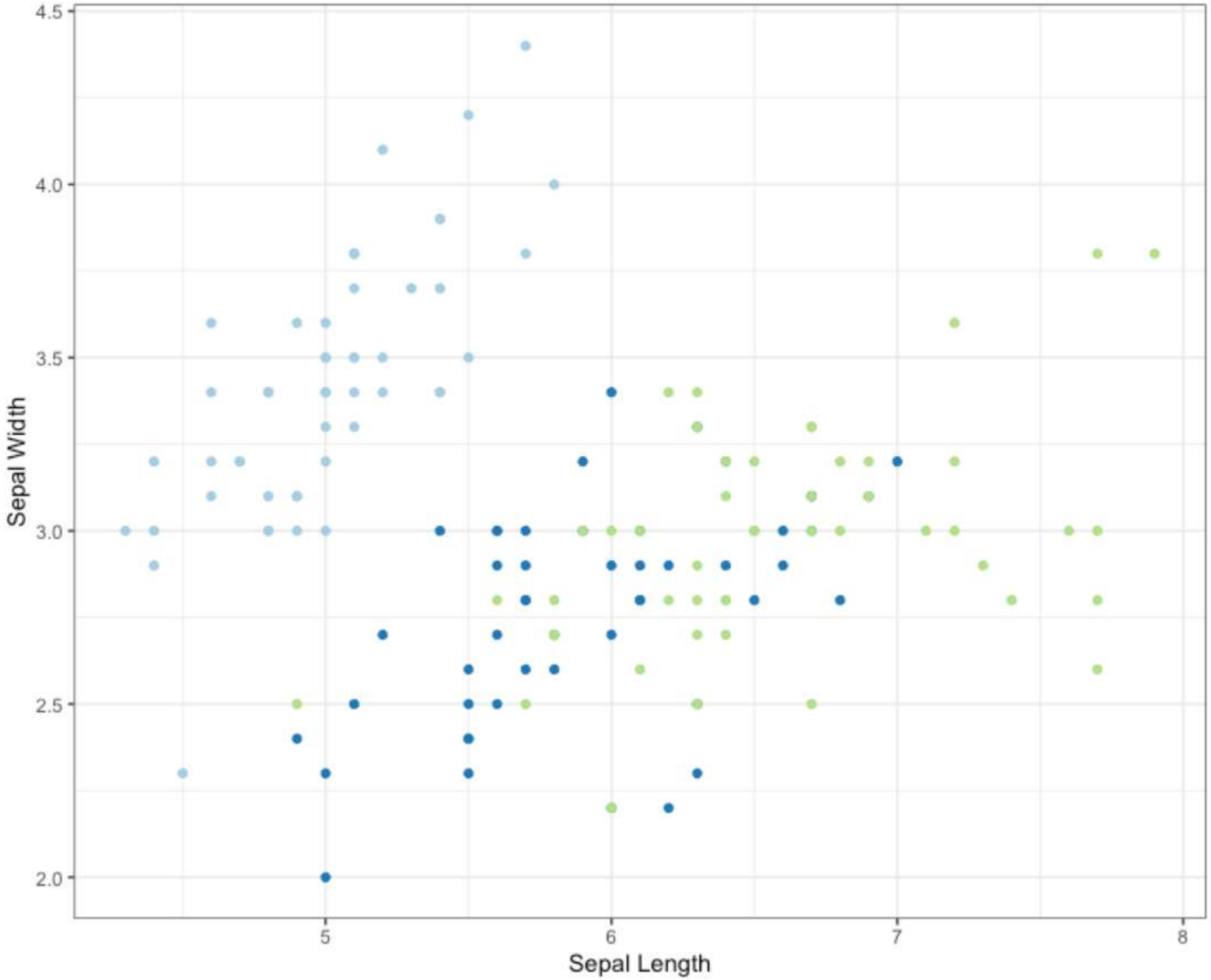
Cleaning up your plot

Adjusting the axes
+scale_x_continuous(“title of x axis”)
+ scale_y_continuous(“title of y axis”)

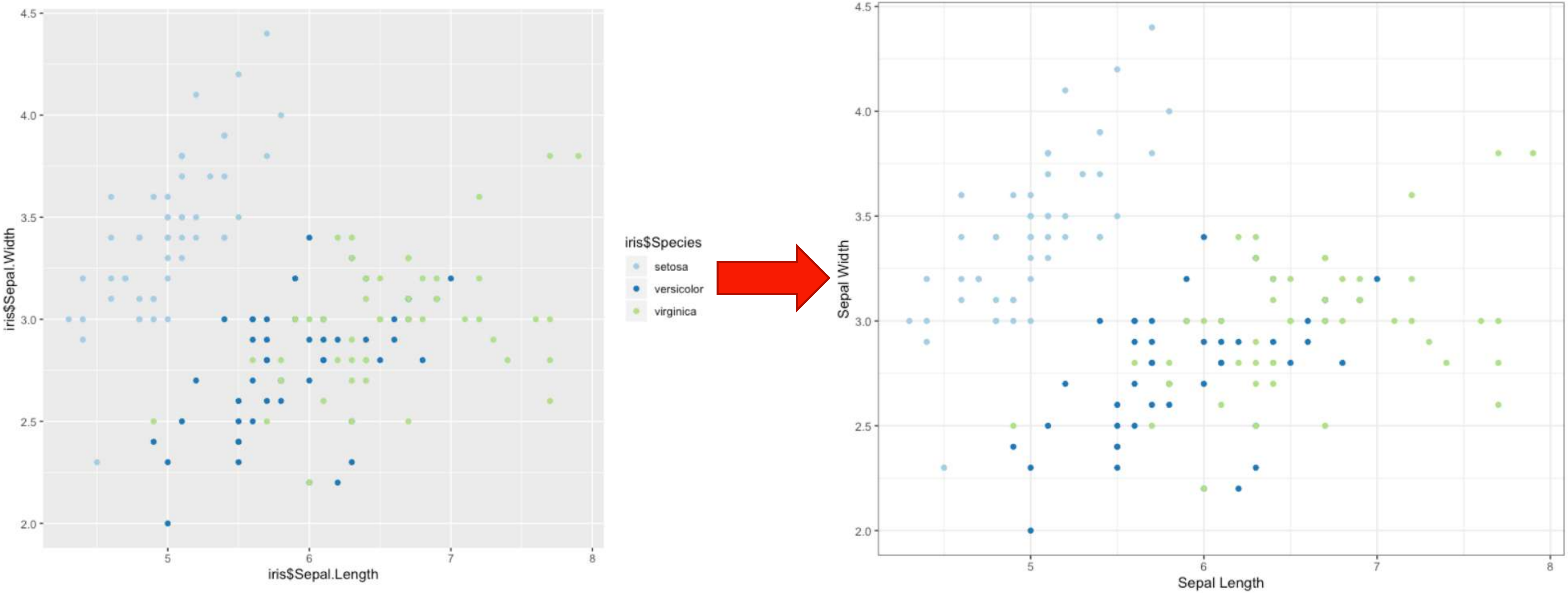


Cleaning up your plot

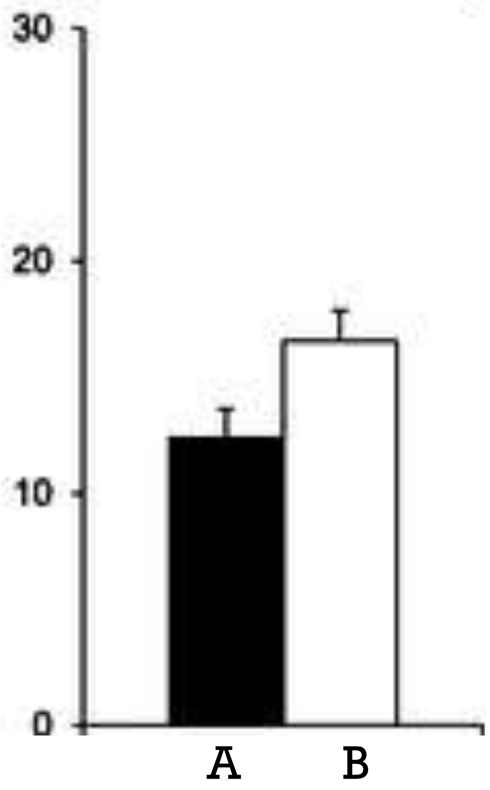
Removing the legend



Before and After of plot



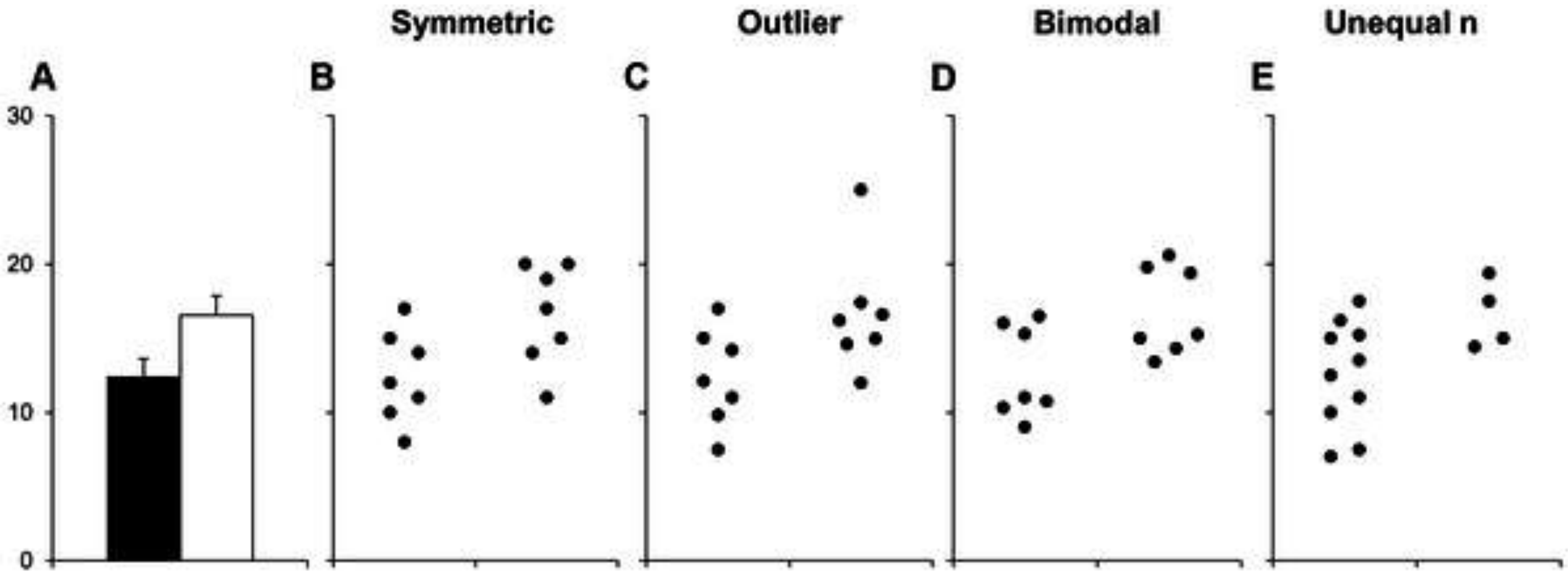
What does this figure communicate about Group A and Group B?



Weissgerber TL, Milic NM, Winham SJ, Garovic VD (2015) Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. PLOS Biology 13(4): e1002128. https://doi.org/10.1371/journal.pbio.1002128
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

PLOS | BIOLOGY
FIFTEENTH ANNIVERSARY

Fig 1. Many different datasets can lead to the same bar graph.

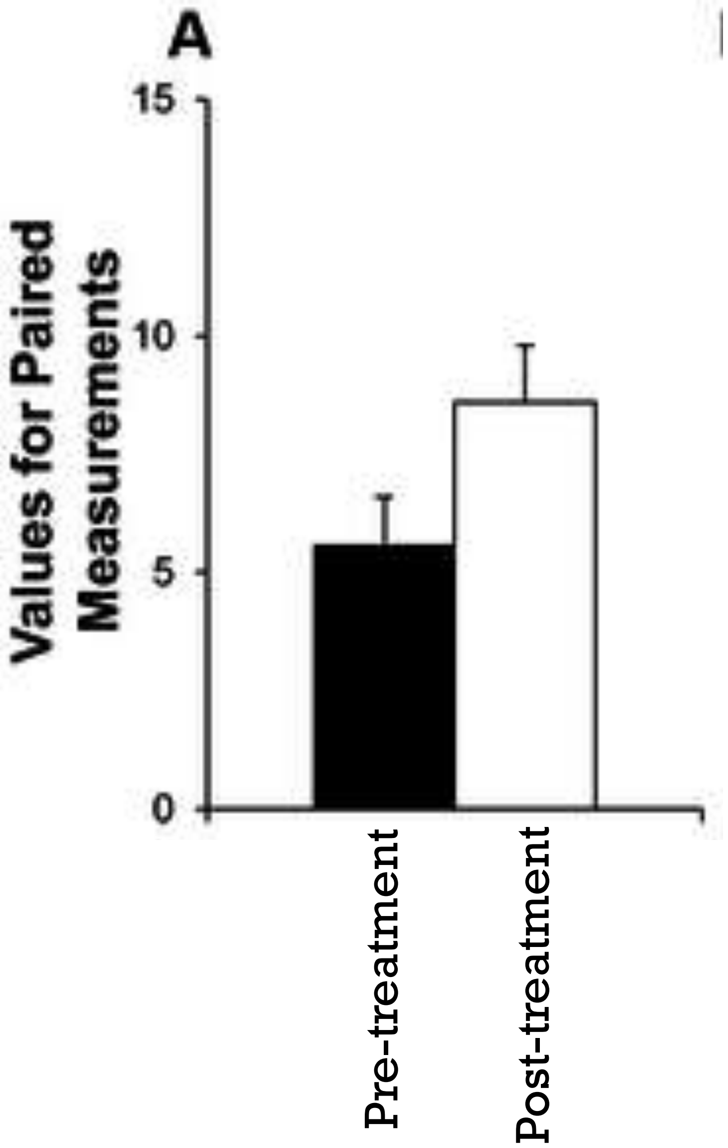


Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

Weissgerber TL, Milic NM, Winham SJ, Garovic VD (2015) Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. PLOS Biology 13(4): e1002128. https://doi.org/10.1371/journal.pbio.1002128
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

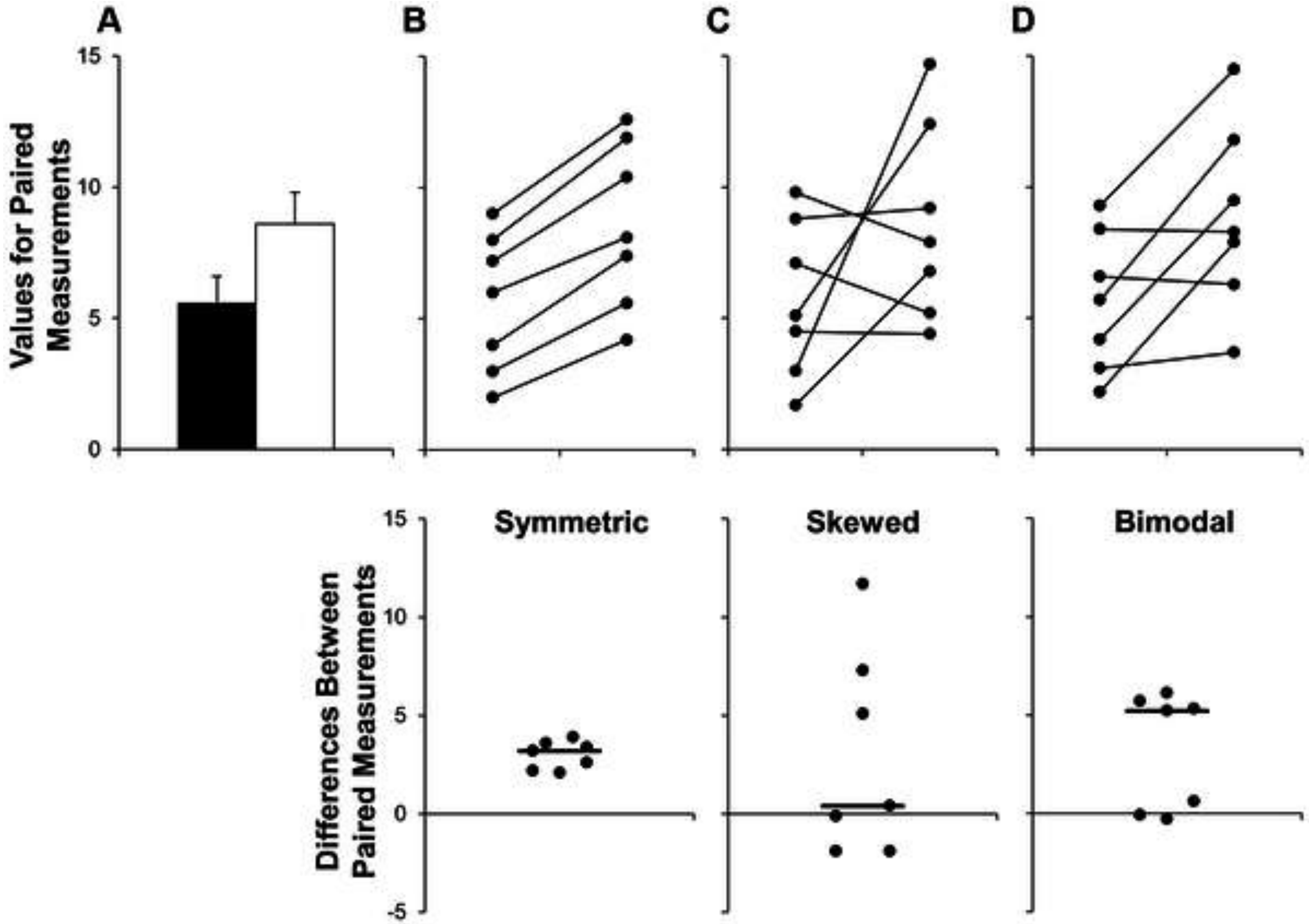
PLOS | BIOLOGY
FIFTEENTH ANNIVERSARY

Fig 2. Additional problems with using bar graphs to show paired data.



Weissgerber TL, Milic NM, Winham SJ, Garovic VD (2015) Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. PLOS Biology 13(4): e1002128. https://doi.org/10.1371/journal.pbio.1002128
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

Fig 2. Additional problems with using bar graphs to show paired data.



Weissgerber TL, Milic NM, Winham SJ, Garovic VD (2015) Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. PLOS Biology 13(4): e1002128. https://doi.org/10.1371/journal.pbio.1002128
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

What about this one?

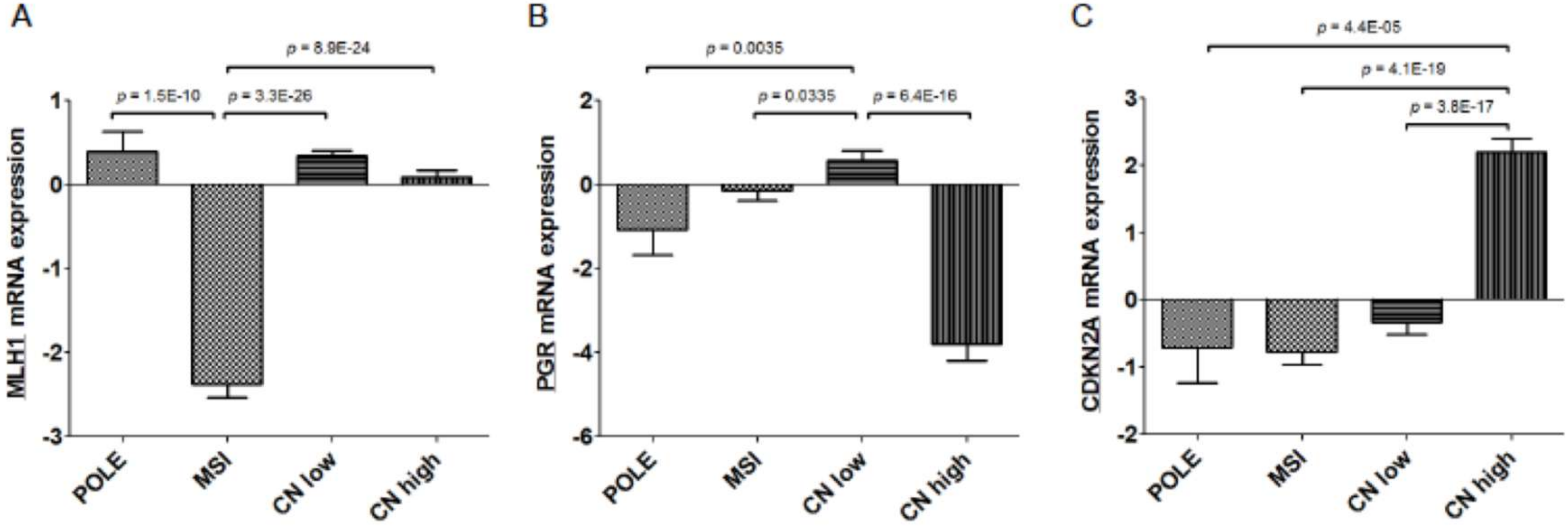


Figure S4.2 Gene expression across integrated subtypes. (A) *MLH1* mRNA expression is significantly lower in the MSI cluster. (B) *PGR* mRNA expression is significantly higher in the CN low cluster. (C) *CDKN2A* mRNA expression is significantly higher in the CN high cluster.

Integrated genomic characterization of endometrial carcinoma, Levine, D.A., 2013

Takeaway message:

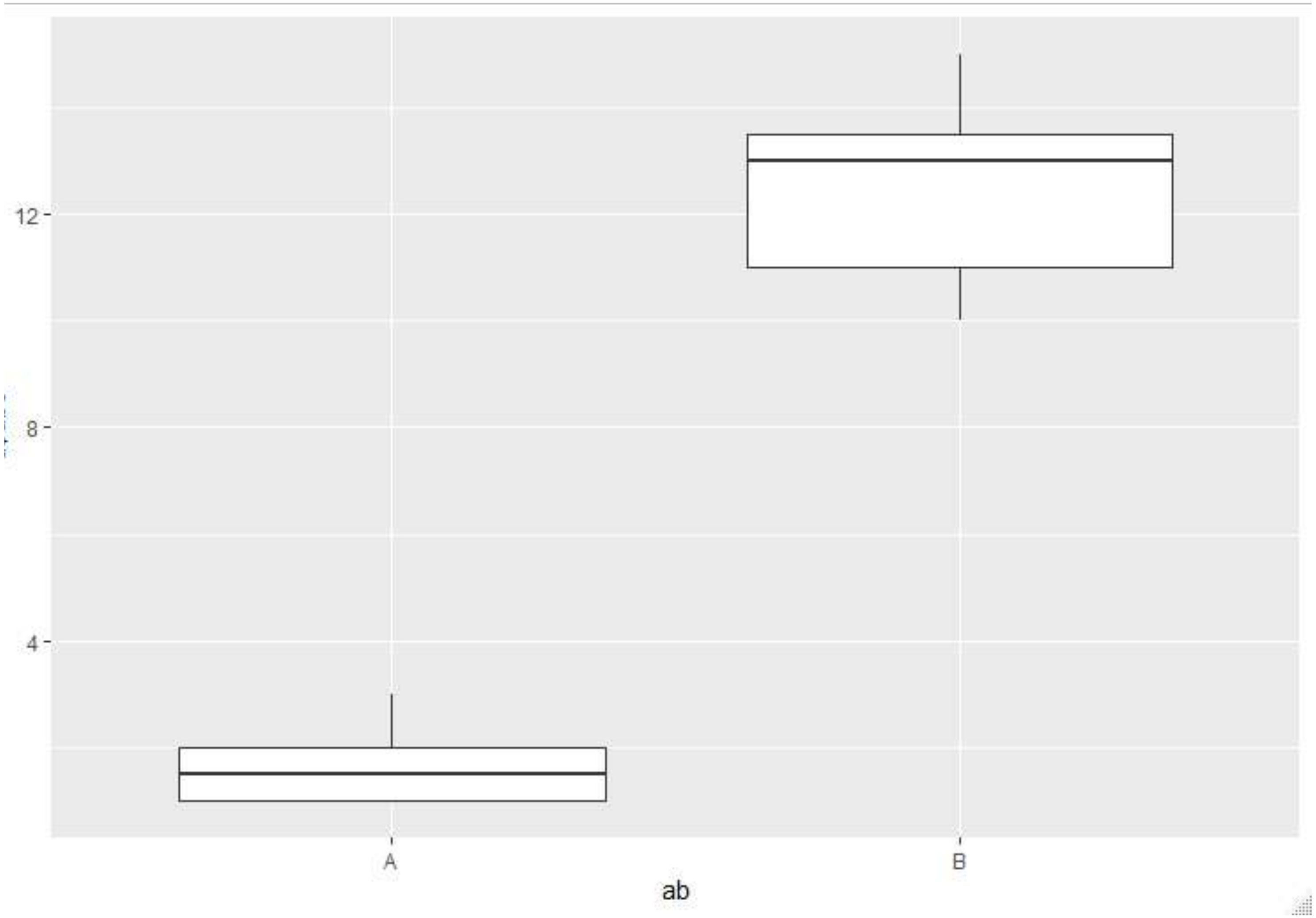
DON’T hide your data.

Avoid inappropriate barplot use.

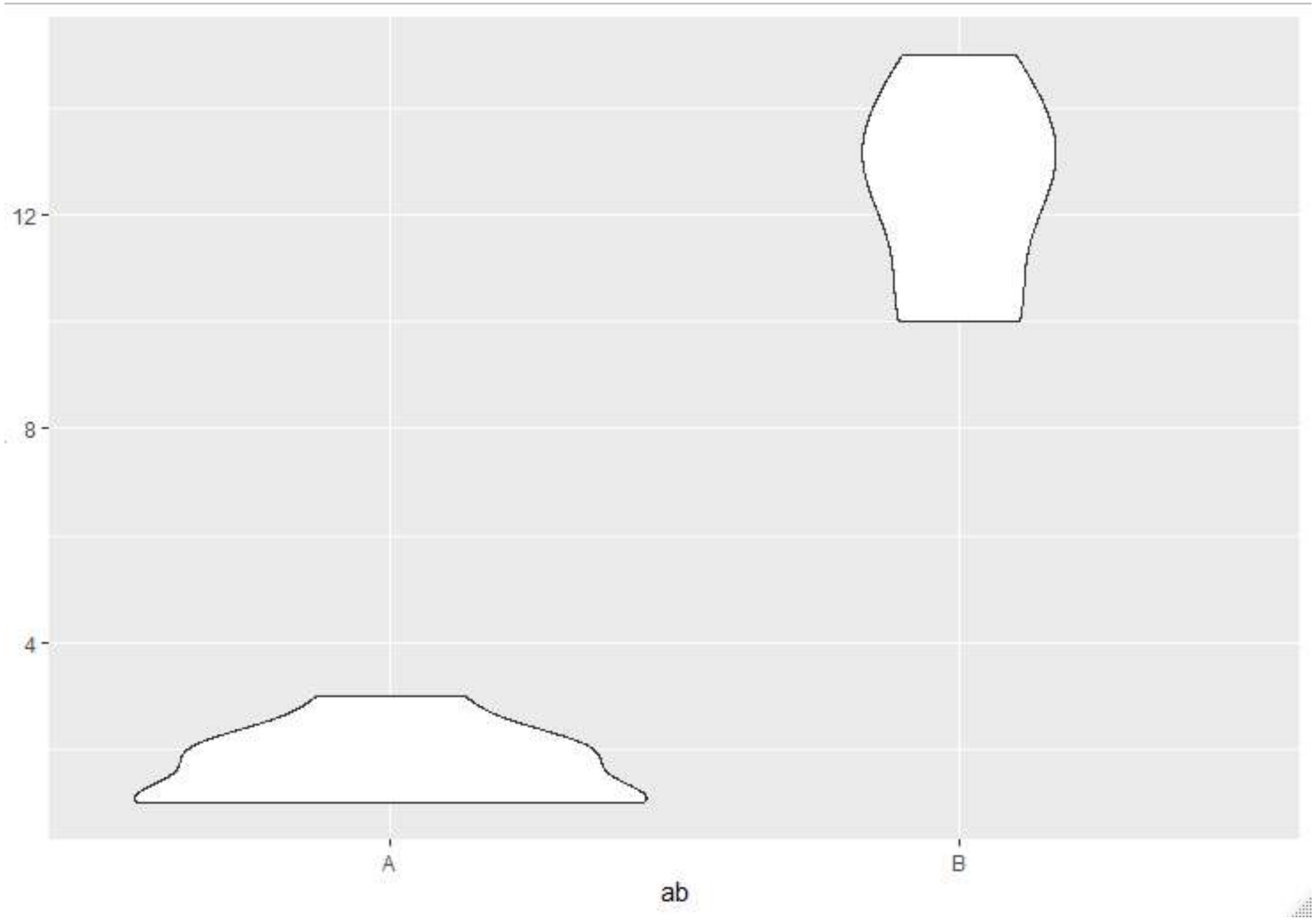
If you don’t need to use barplots, DON’T.

A Return to Making Boxplots:

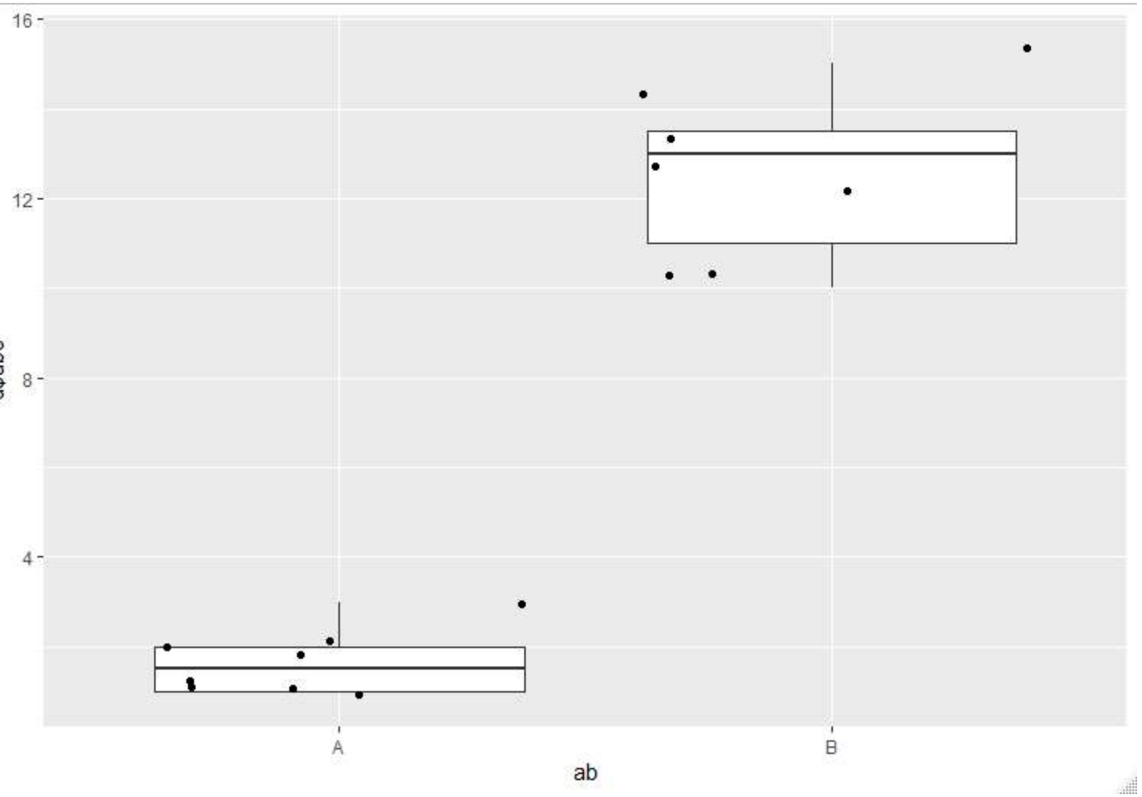
Given what we just discussed, what potential issues exist with boxplots like this one?

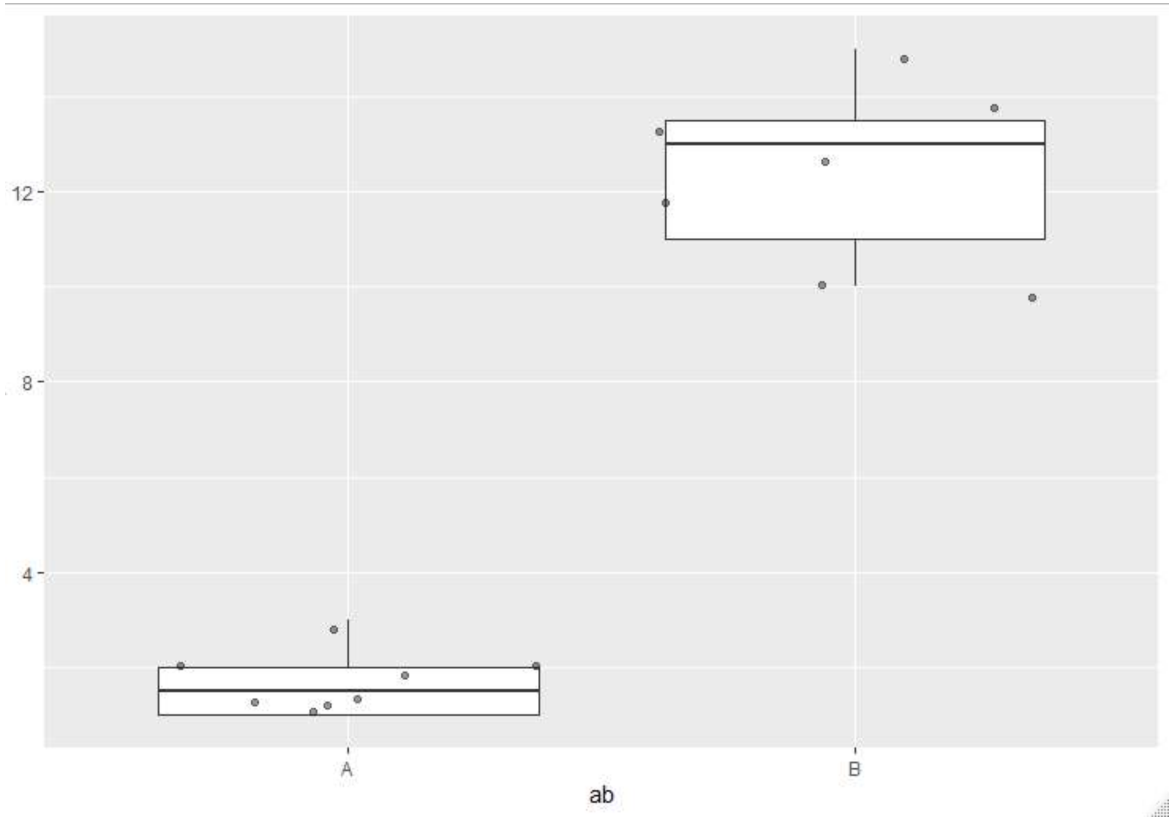


Beyond Basic Boxplots: the Violin Plot

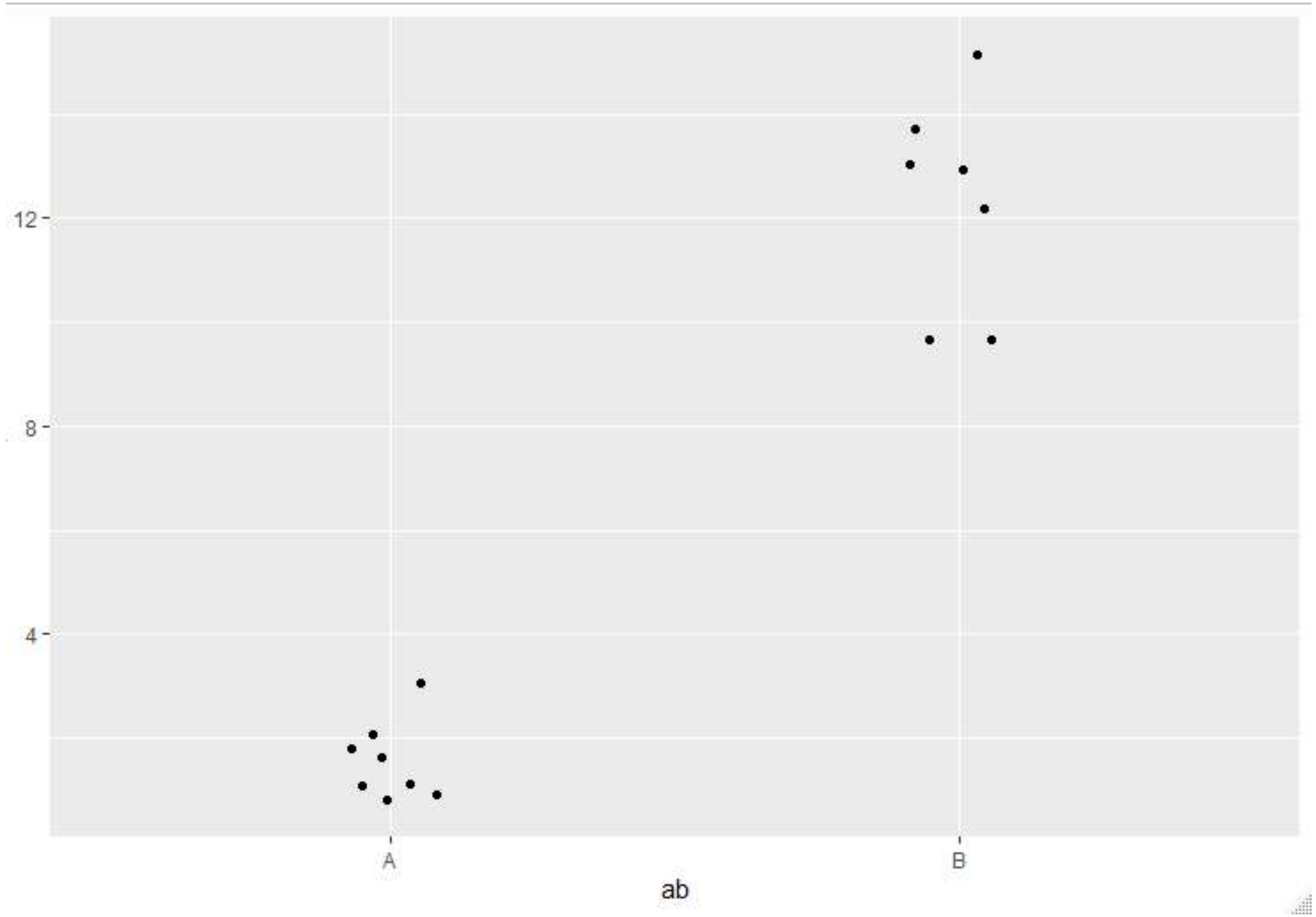


Beyond Basic Boxplots: Overlaying Jitter Points

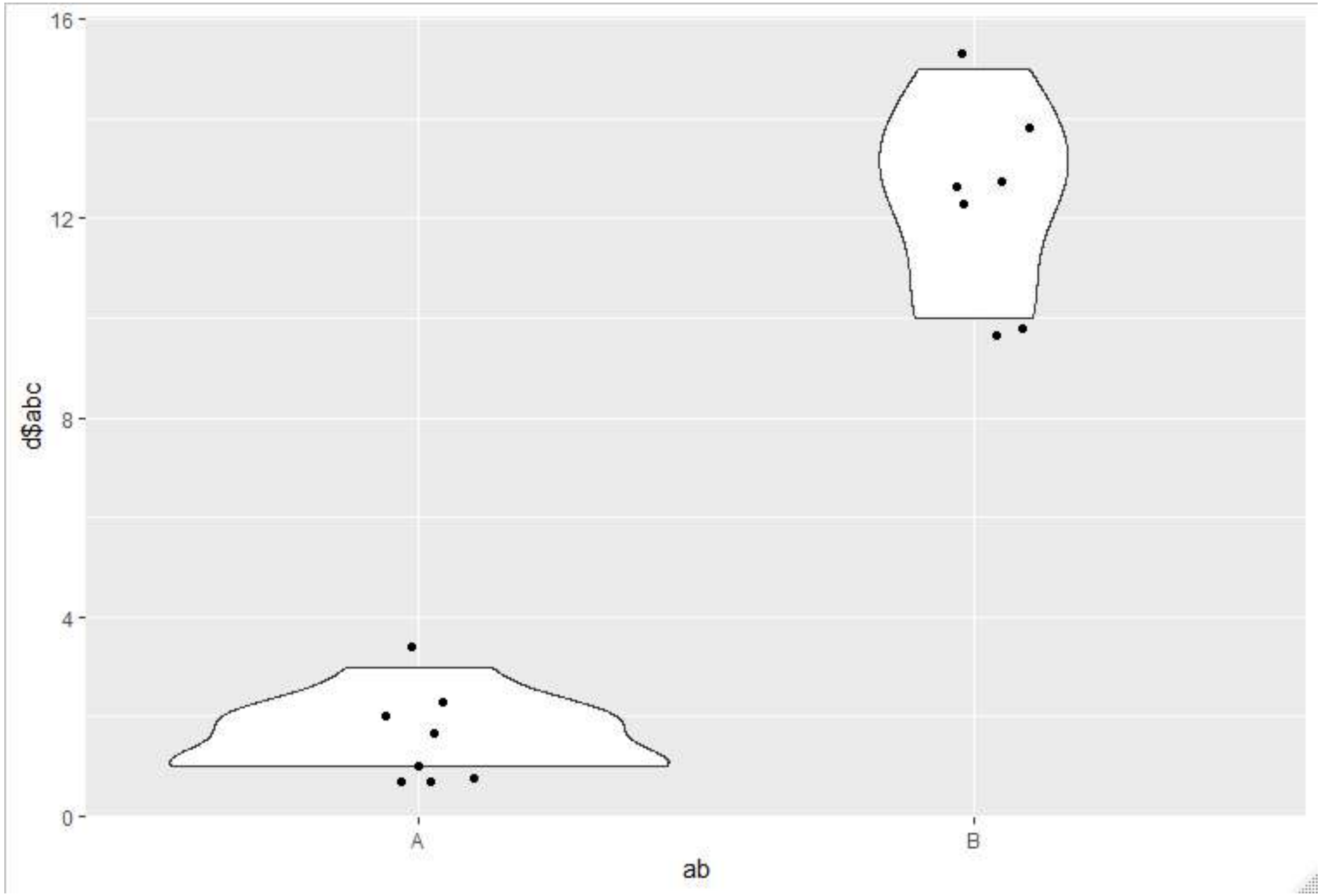




Beyond Basic Boxplots: Strip Charts



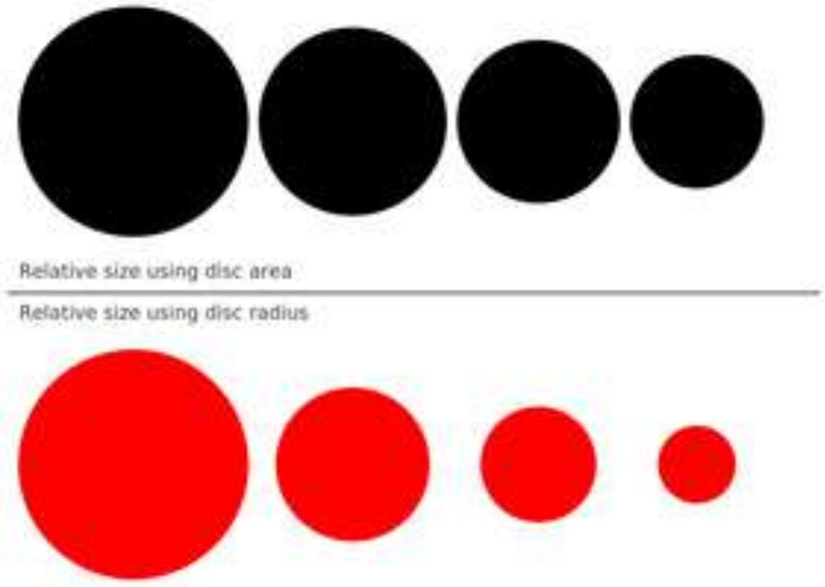
Beyond Basic Boxplots: Violin Plots + Jitter Points



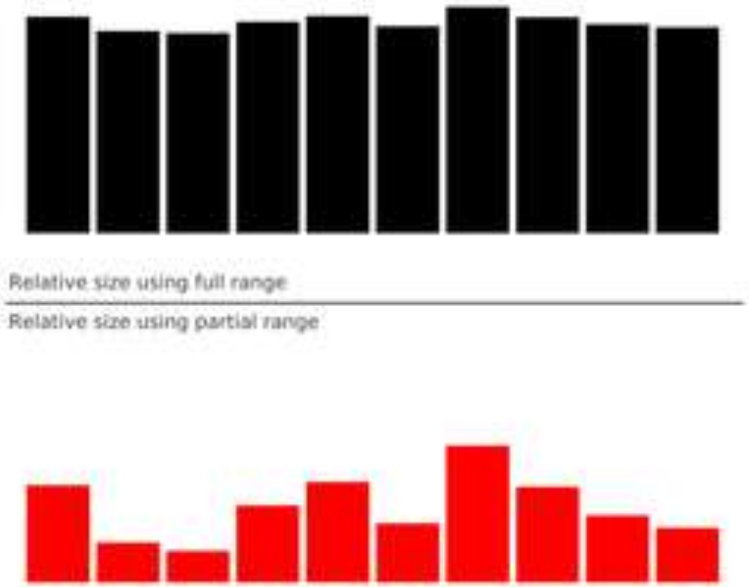
Takeaway message:

- Show off your data, you worked hard to get it!
- Invite your readers to analyze and engage with your material by being transparent
- Reinforce the validity of your findings

Do not mislead the reader.


Relative size using disc area
Relative size using disc radius

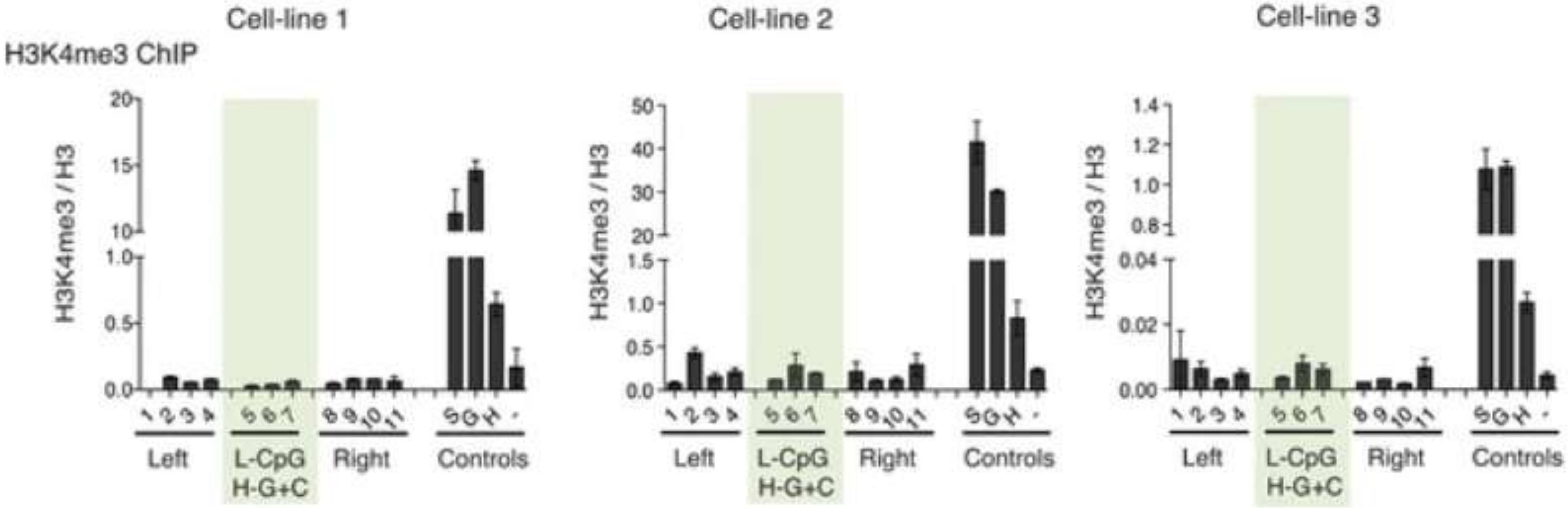
Using Disc Area vs Disc Radius



Using Full range vs Partial Range

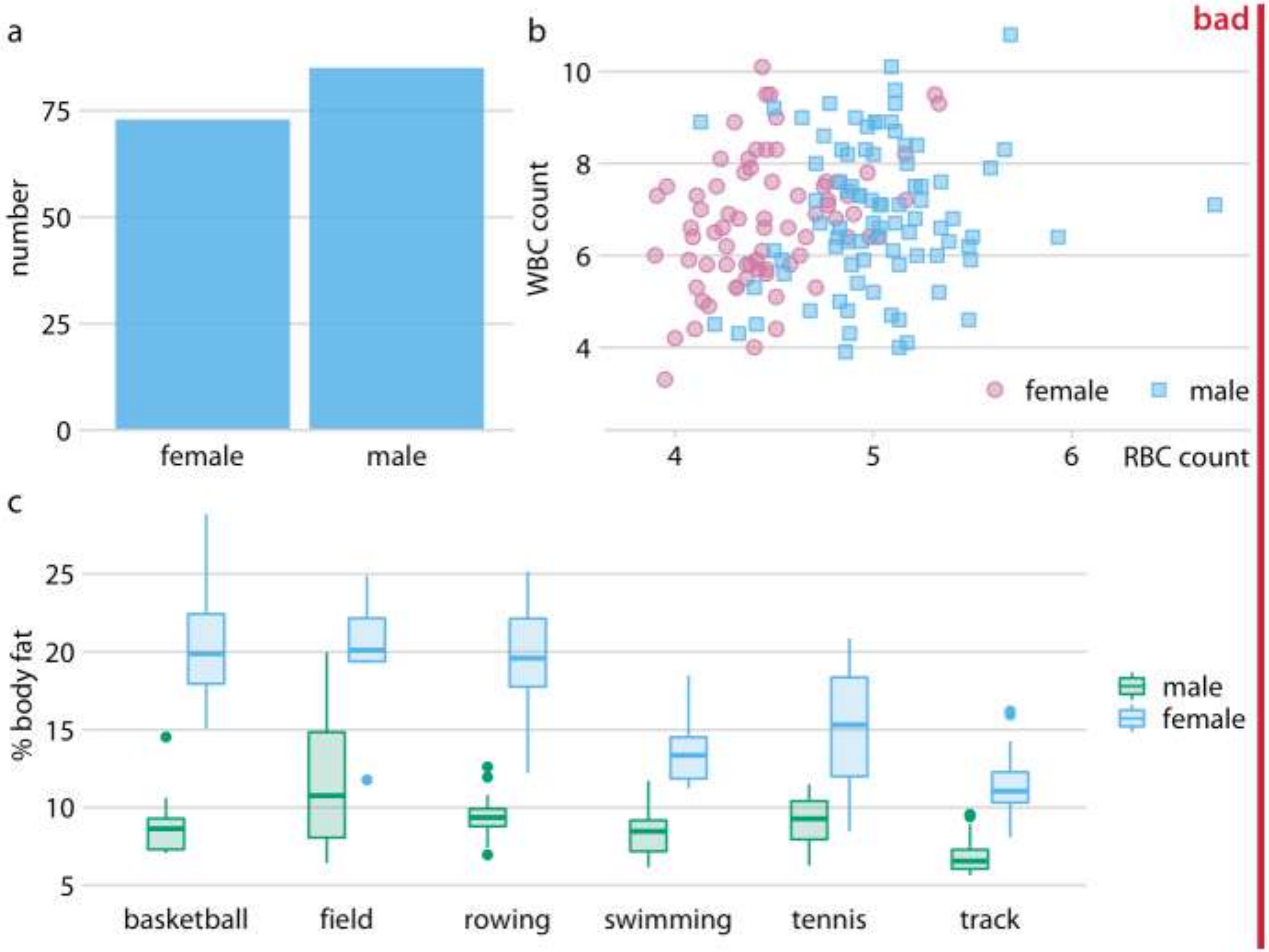
Rougier NP, Droettboom M, Bourne PE (2014) Ten Simple Rules for Better Figures. PLOS Computational Biology 10(9): e1003833. https://doi.org/10.1371/journal.pcbi.1003833
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833

Example from literature: what issues do you identify with this figure?



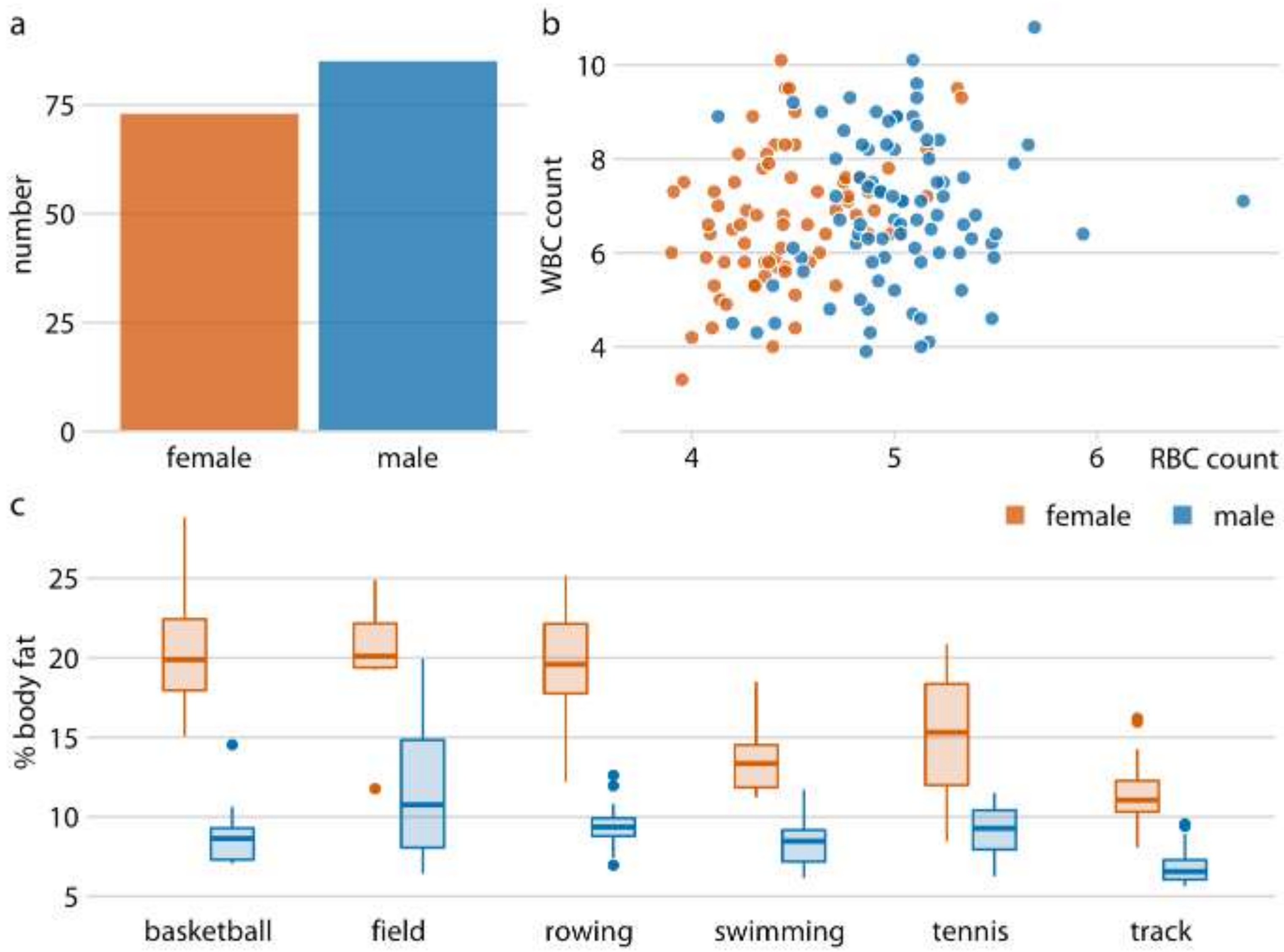
Synthetic CpG islands reveal DNA sequence determinants of chromatin structure, Wachter E. et al, 2014

What is this figure trying to convey? Is it easy to process and understand?



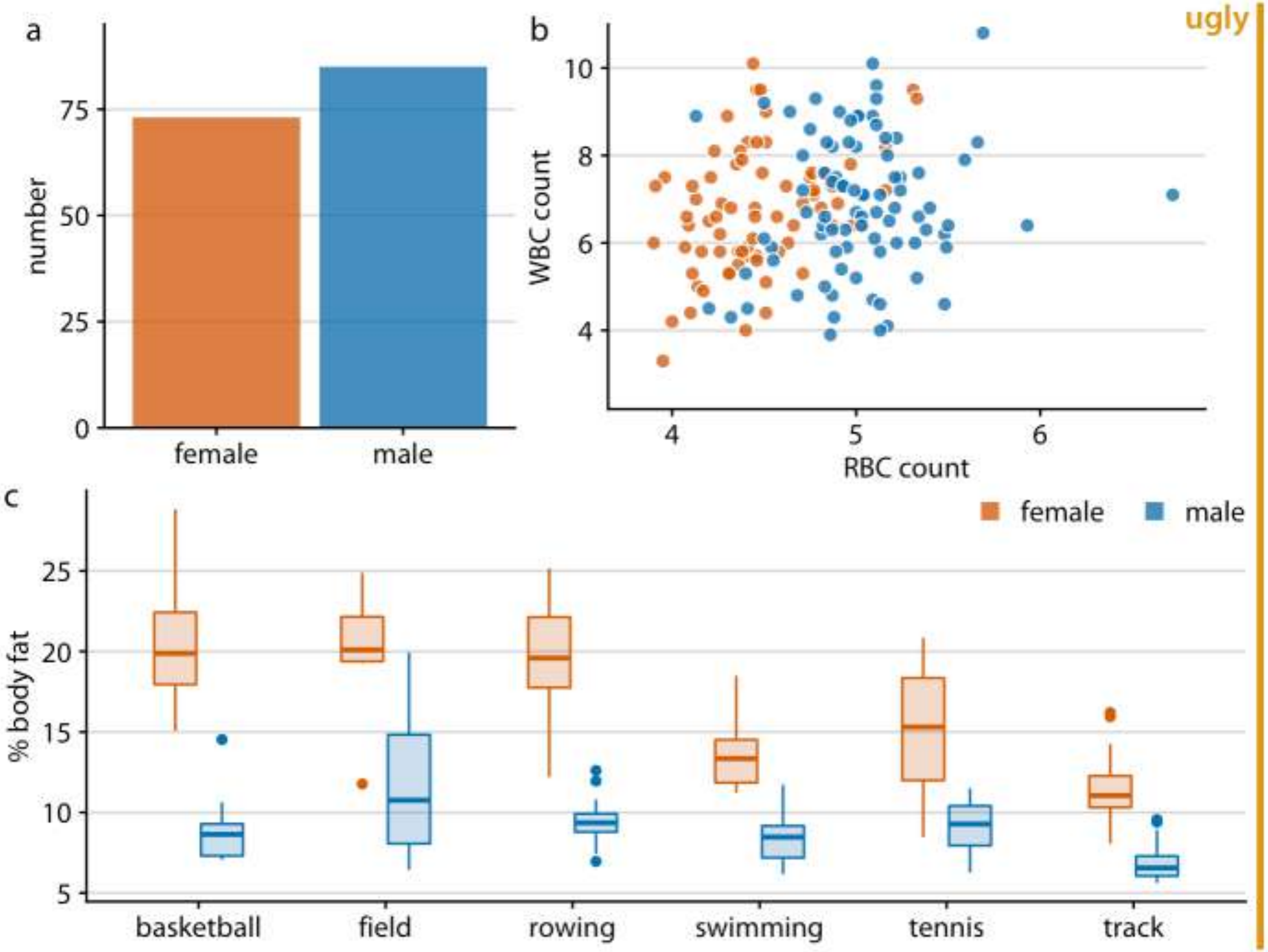
https://serialmentor.com/dataviz/multi-panel-figures.html

Having a **consistent visual language** is important for conveying a **strong message**



Only one legend needed now

When possible, make sure that figures are well aligned!



Remember:

Good Data Visualization =
Good Communication =
Transparency =
Good Science!!

Exercise 6:

Let’s revisit the plots we made in Exercise 5. Given everything that was covered in the presentation, let’s discuss what is good about those plots and how they can be improved!

Exercise 7:

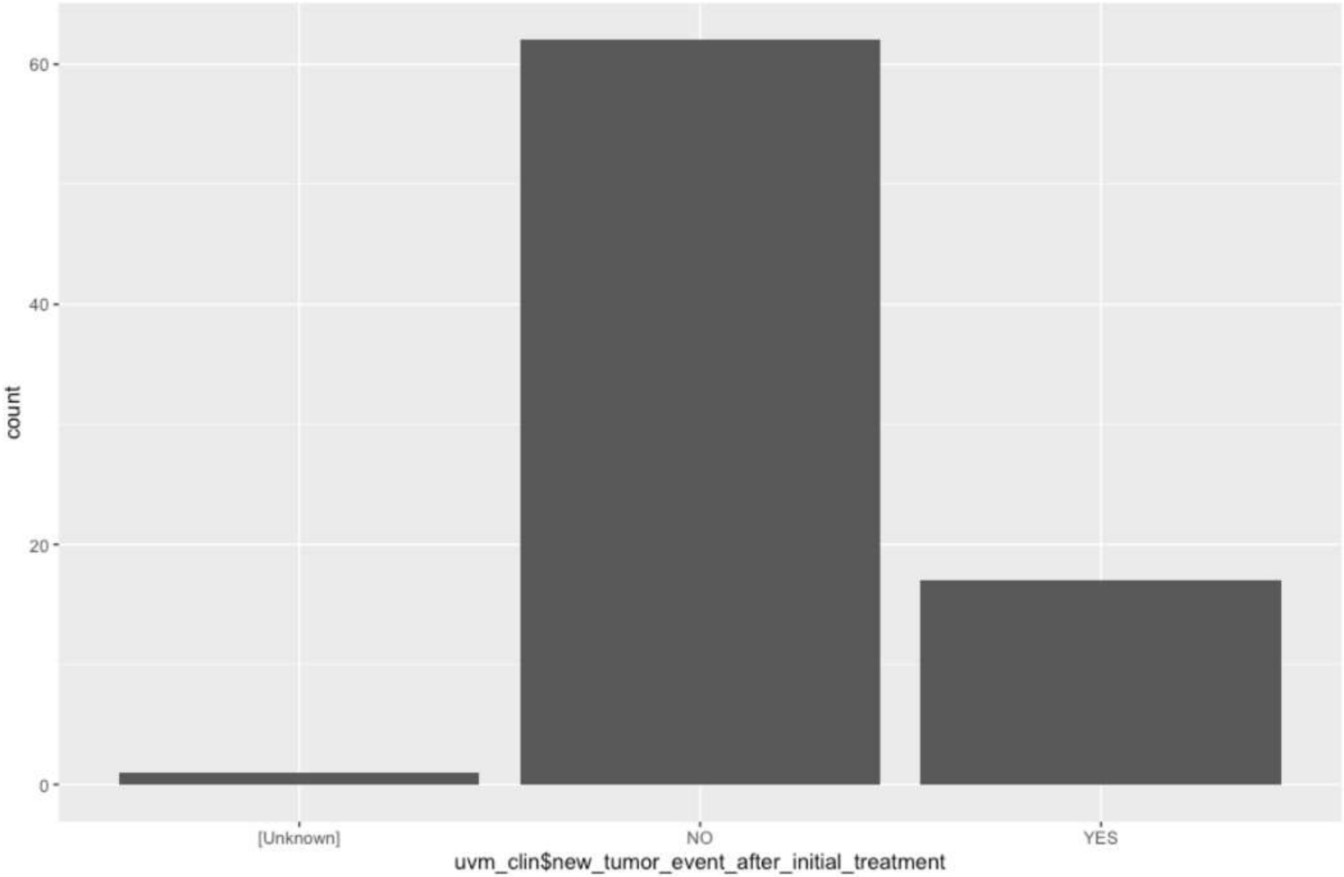
Take the plot you made in Exercise 5 and make adjustments and optimizations based on the principles of data visualization and figure design that we saw in the Presentation section.

Post your before and after in the Google Doc!

Exploration

Facet_grid

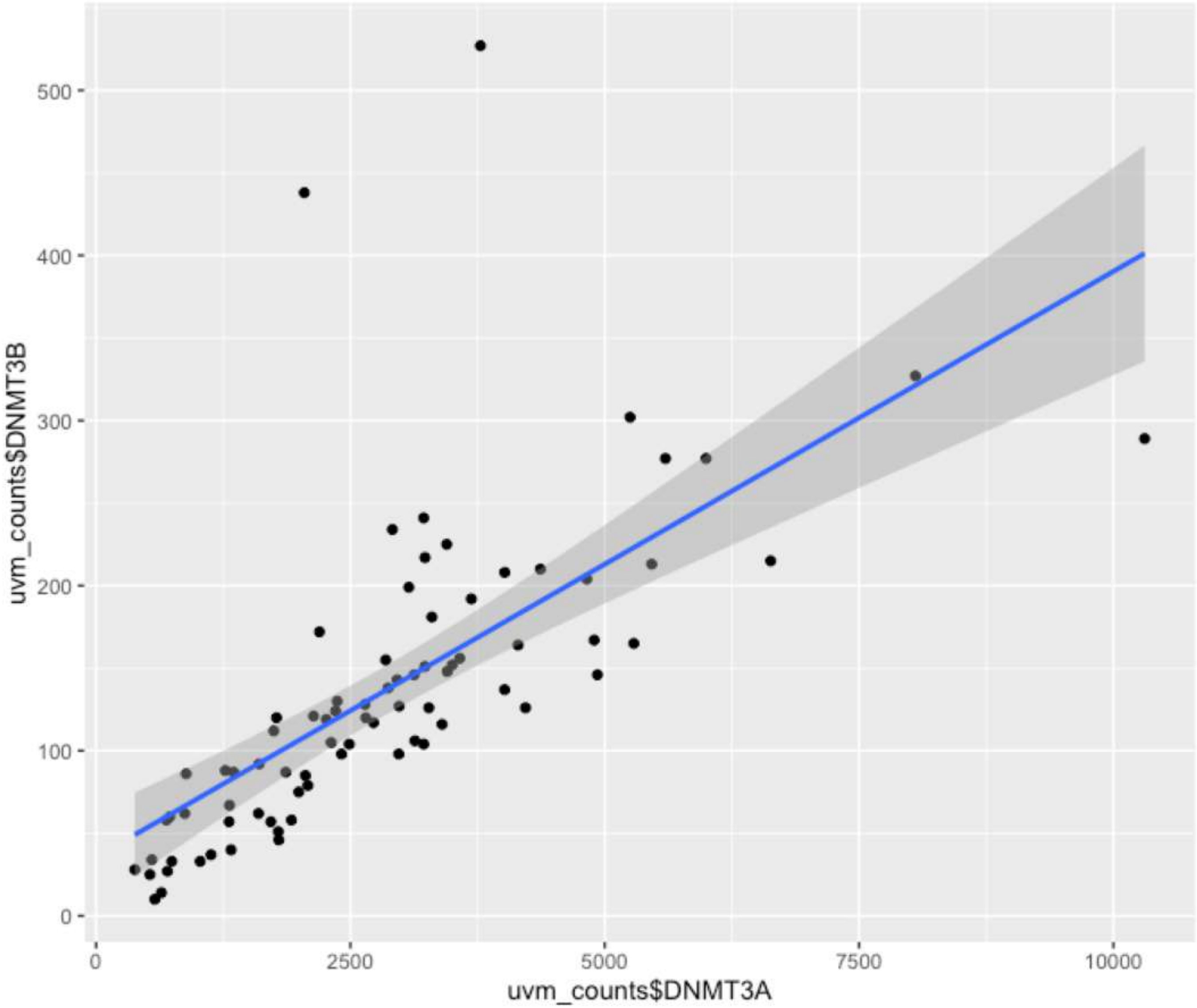
How can we easily check whether tumor recurrence varies across eye colours, or if it is consistent across eye colors?



Linear regression lines

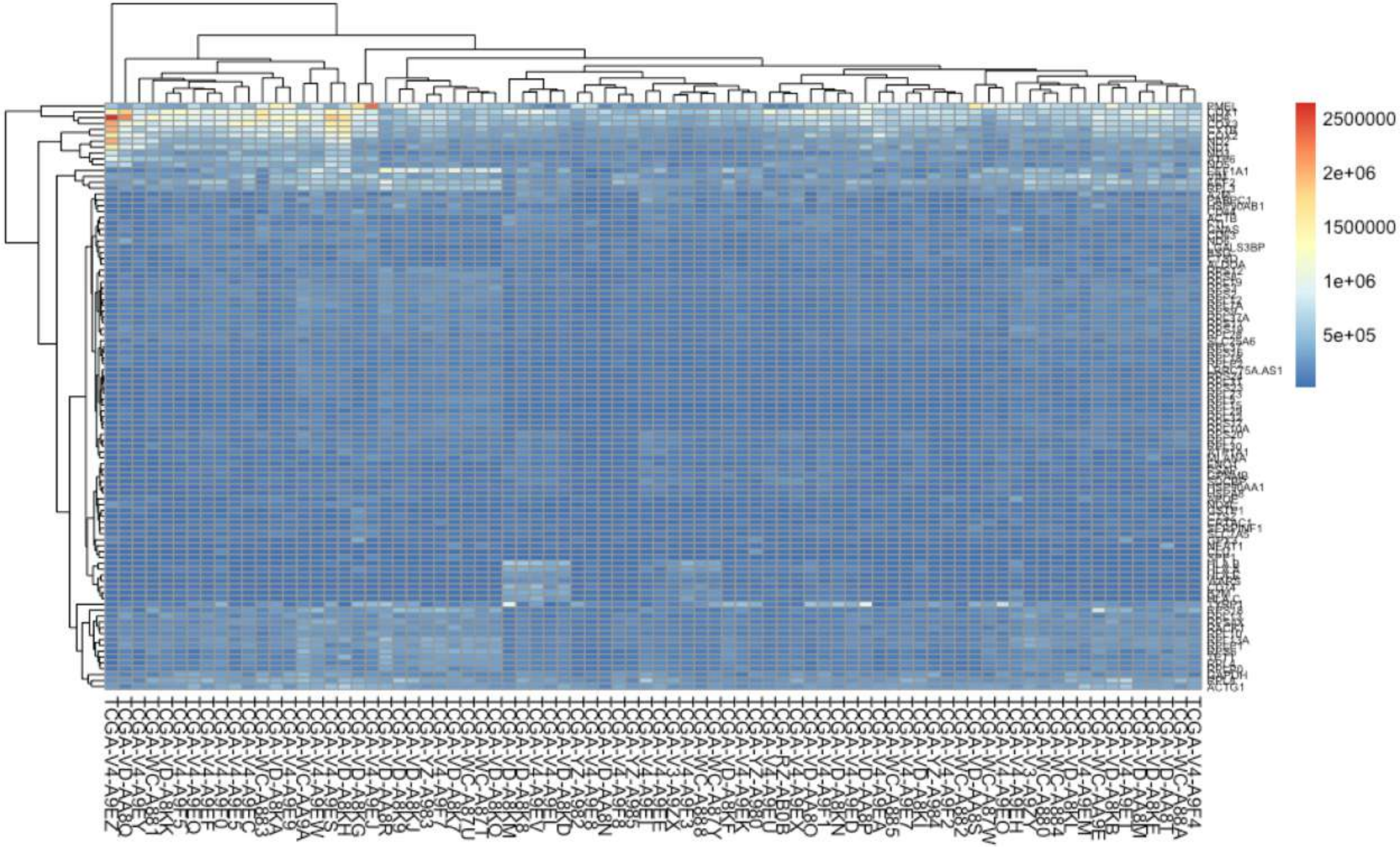
To add a linear regression line to your scatter plot:

```
+ stat_smooth(method = lm)
```



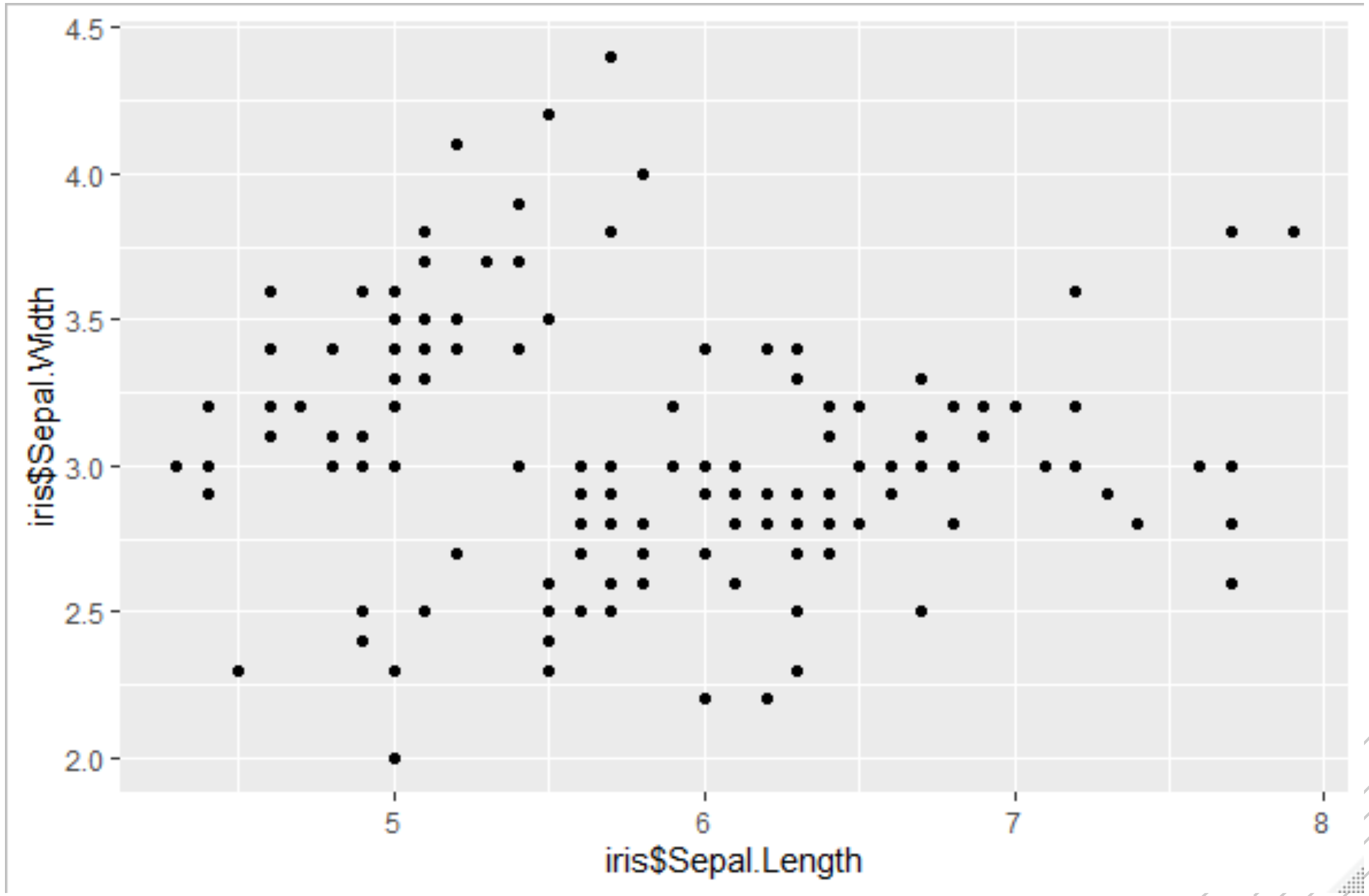
Heatmap

- There are several packages that can help you generate a heatmap, each with their own strengths
- As an example, here is a heatmap made using pheatmap:

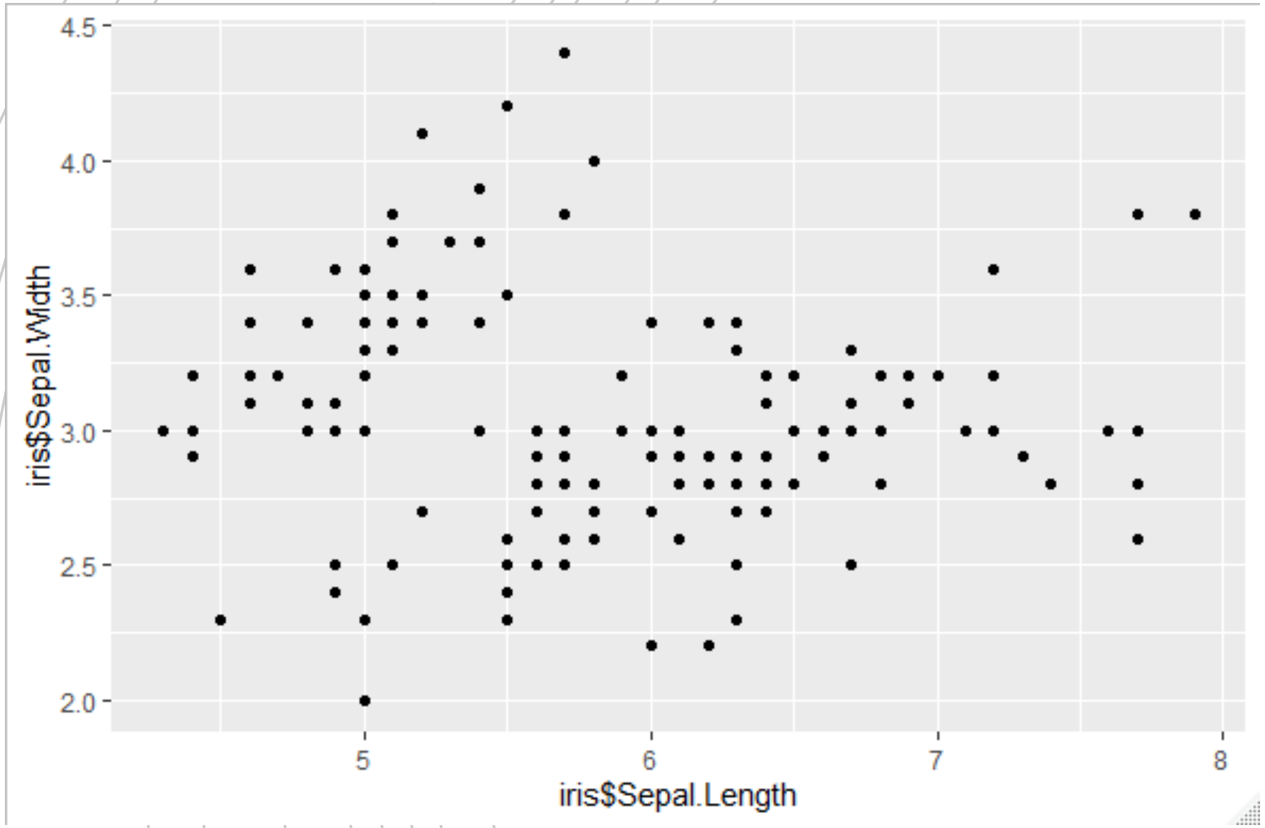


Interactive Plotting

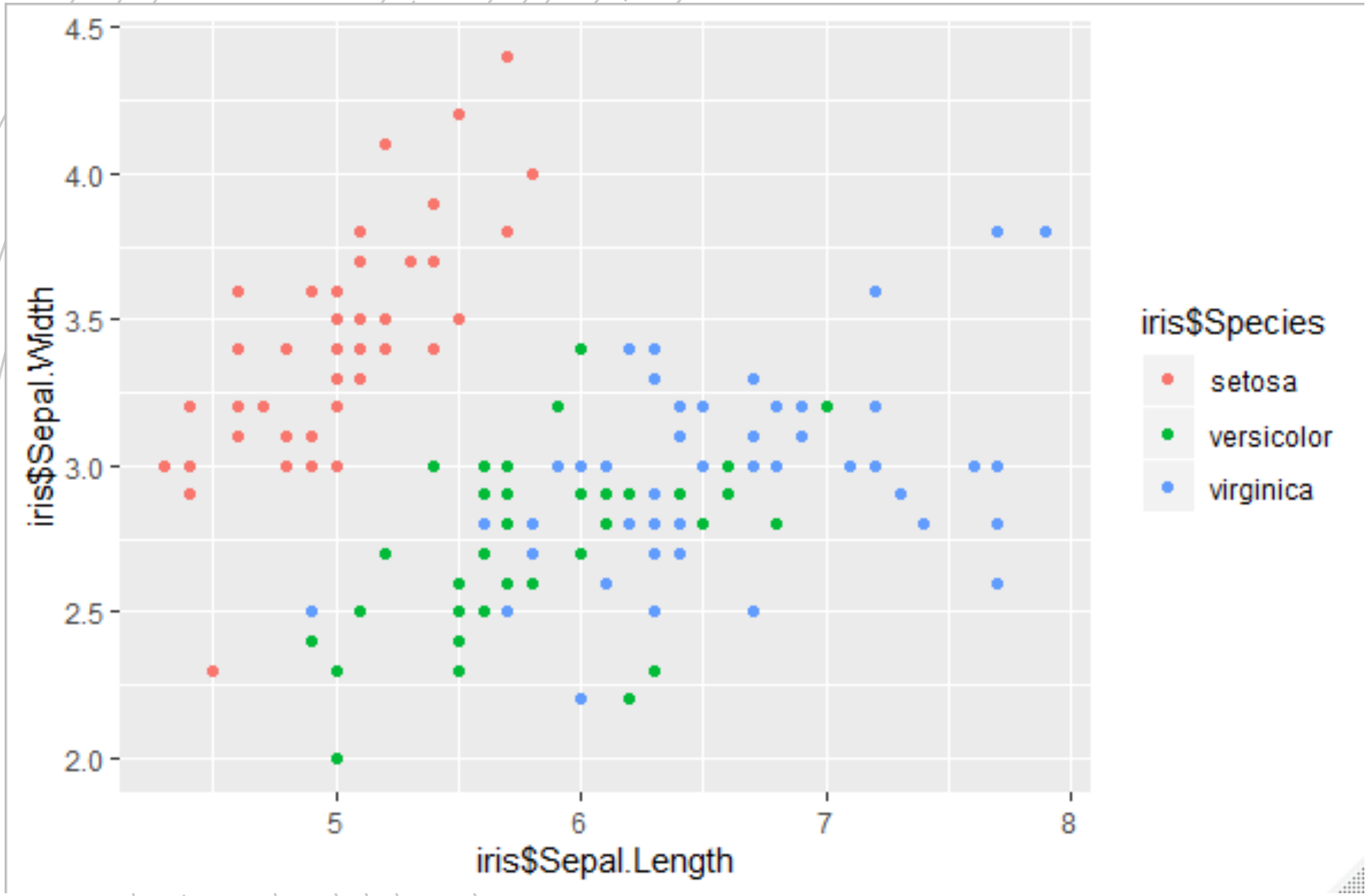
- Helps provide a more intuitive approach to data exploration, adding layers of information to the plot.
- For example:



What can you say about the data here?
What questions do you have about this plot?



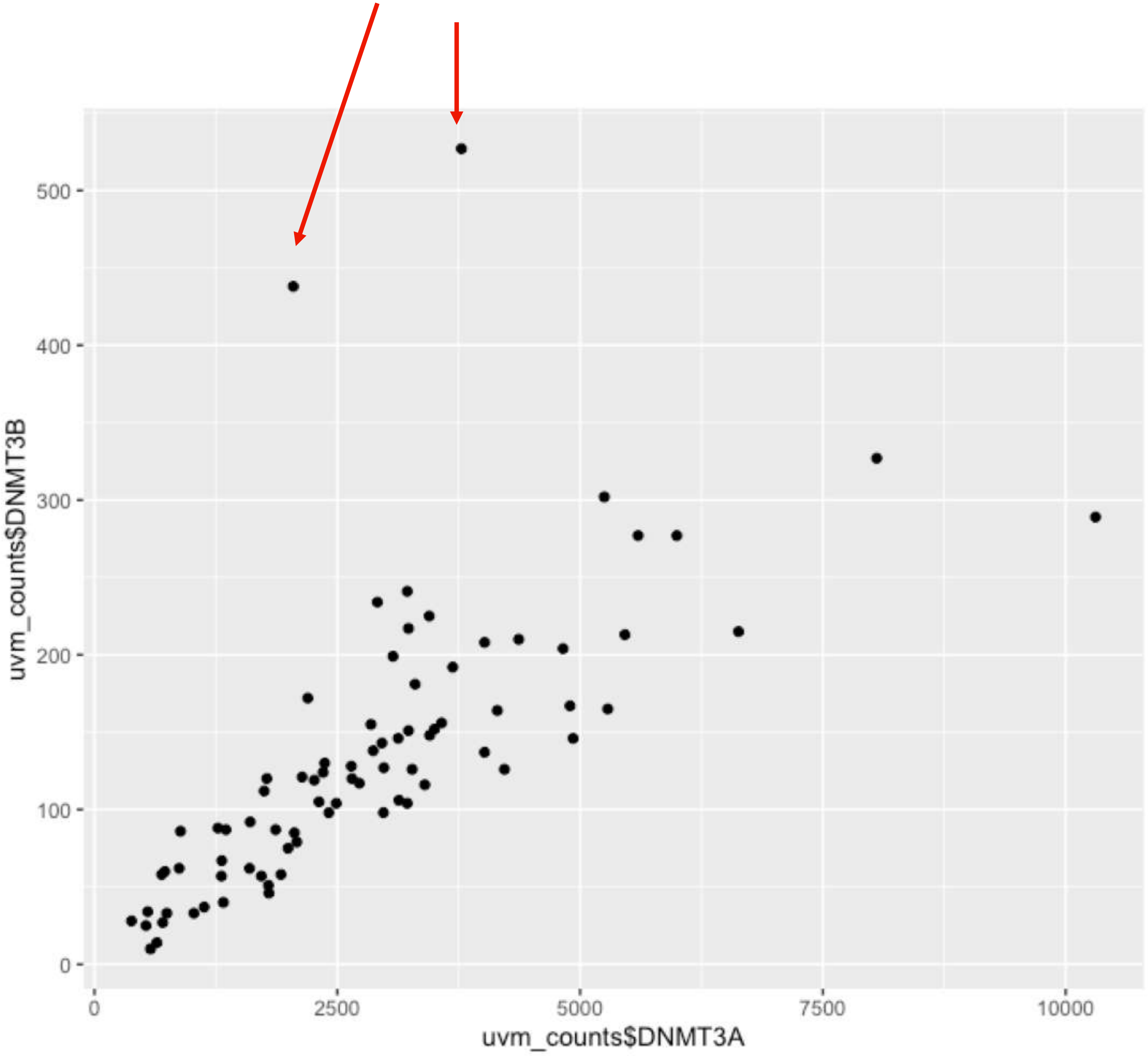
Are there subpopulations here that explain the point distribution?



- Following up on hypotheses via color may be one option
- But limited information is known on each individual point

Revisiting our scatter plot from the beginning…

Which patients are these? How do we find them?



Finding the outliers in the data…

1. We can look through the dataframe for samples that have expression levels that match up with the coordinates we see for the outliers in the scatterplot
2. Or we can make the scatterplot interactive and show the sample name on mouse hover

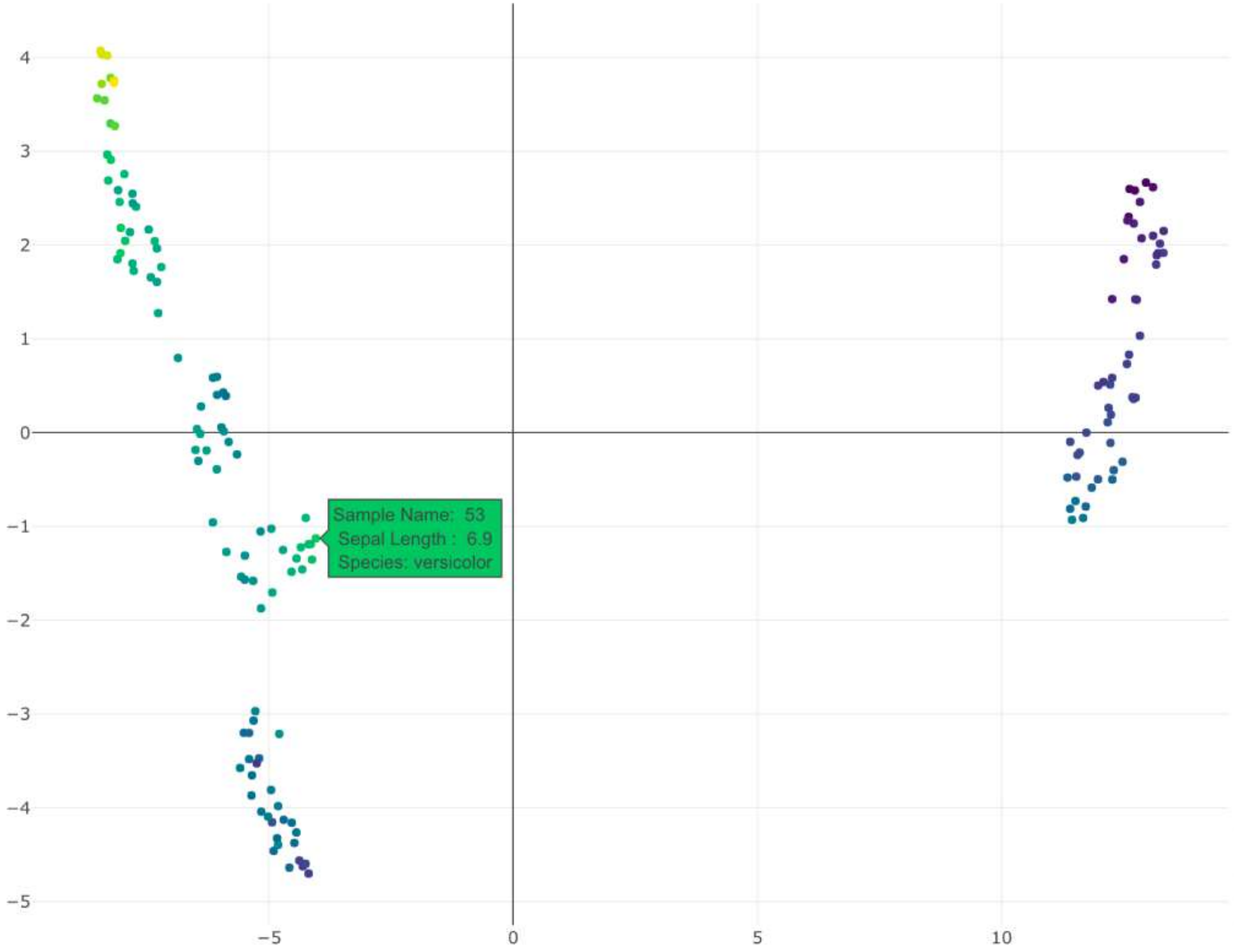
Interactive plots!!

- *ggplotly* is a nice function to get started with interactive plots!
- Let’s install the package that we need to use ggplotly:
- install.packages("plotly")
- library(plotly)
- Steps for basic ggplotly plot generation:

1. Write out the command for the plot you are interested in making, and assign the plot to a variable Ex: a<-ggplot(…)
2. ggplotly(a)

Interactmapper package

- Using the functions included in interactmapper, you can add extra information to the plots resulting from dimension-reduction methods, via color scheme and interactive elements

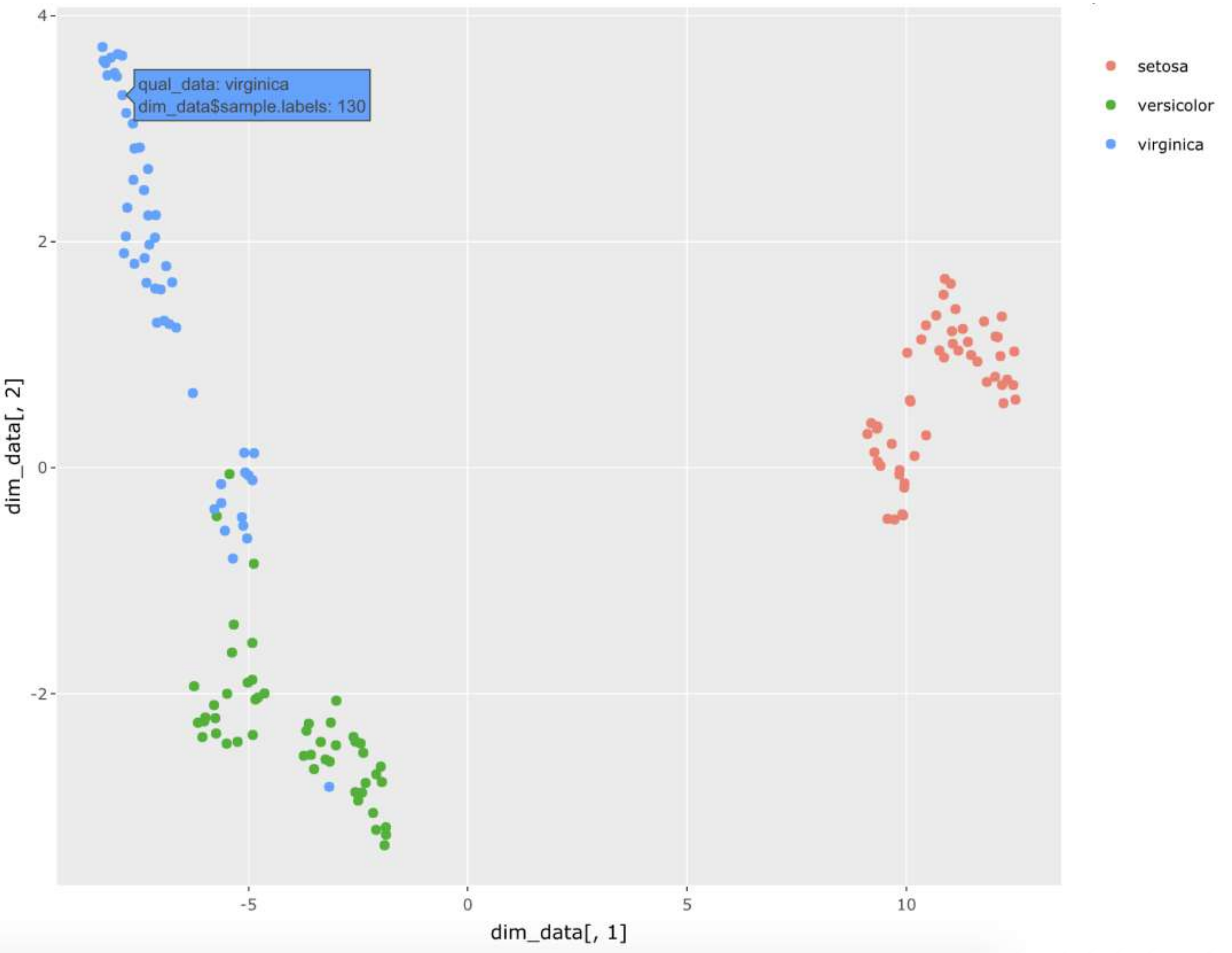


Interactmapper package

- Currently there are three functions available to generate interactive plots in the interactmapper package:
  - interact_qual
  - interact_quant
  - interact_multi

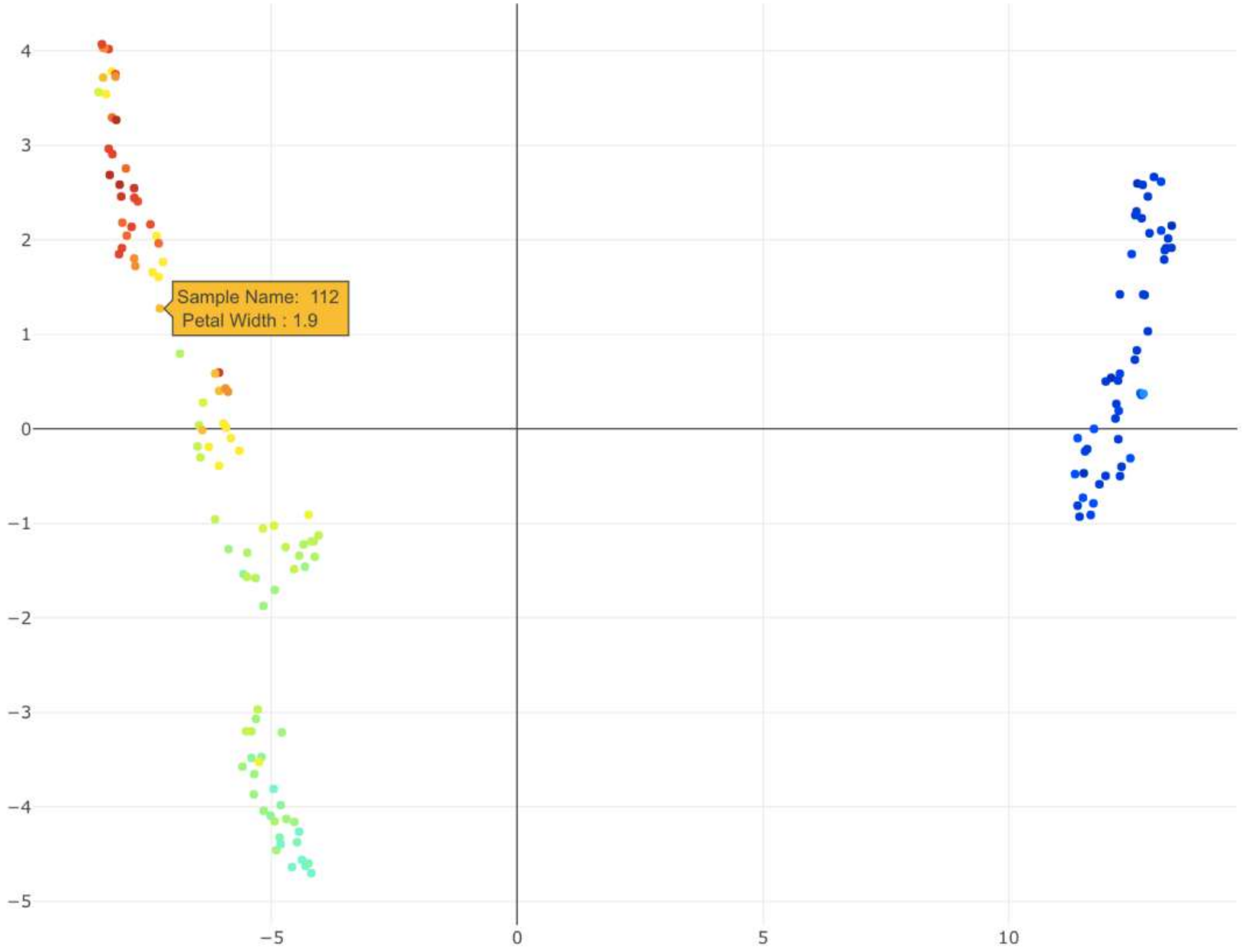
- Ideal for qualitative data
- To use this function: interact_qual(count_data, qual_data, dim_red_meth = c("UMAP", "PCA"))

Interact_qual



- Ideal for quantitative features
- To use:
- interact_quant(count_data, quant_info_name, quant_info, dim_red_meth = c("UMAP", "PCA"), your_palette)

Interact_quant

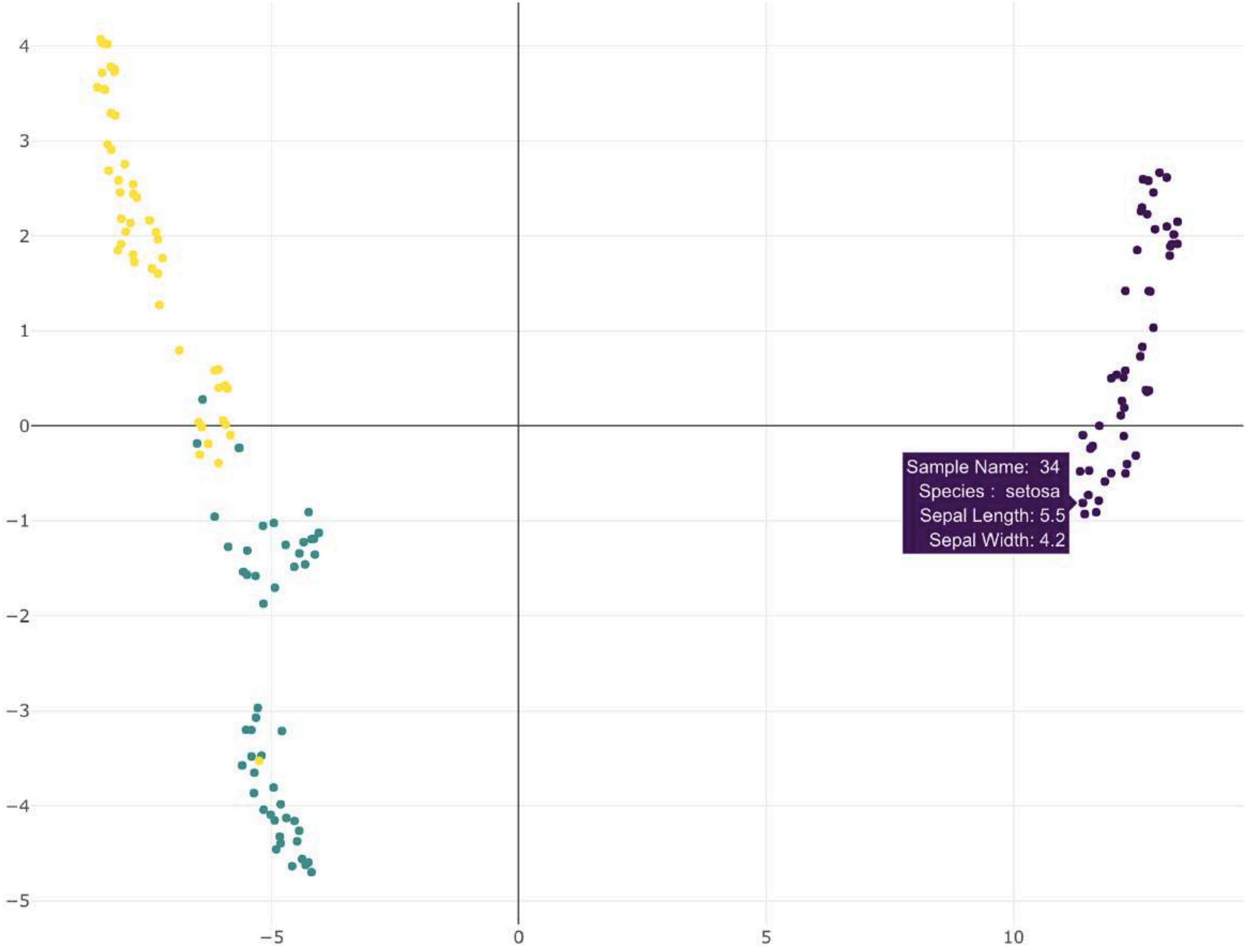


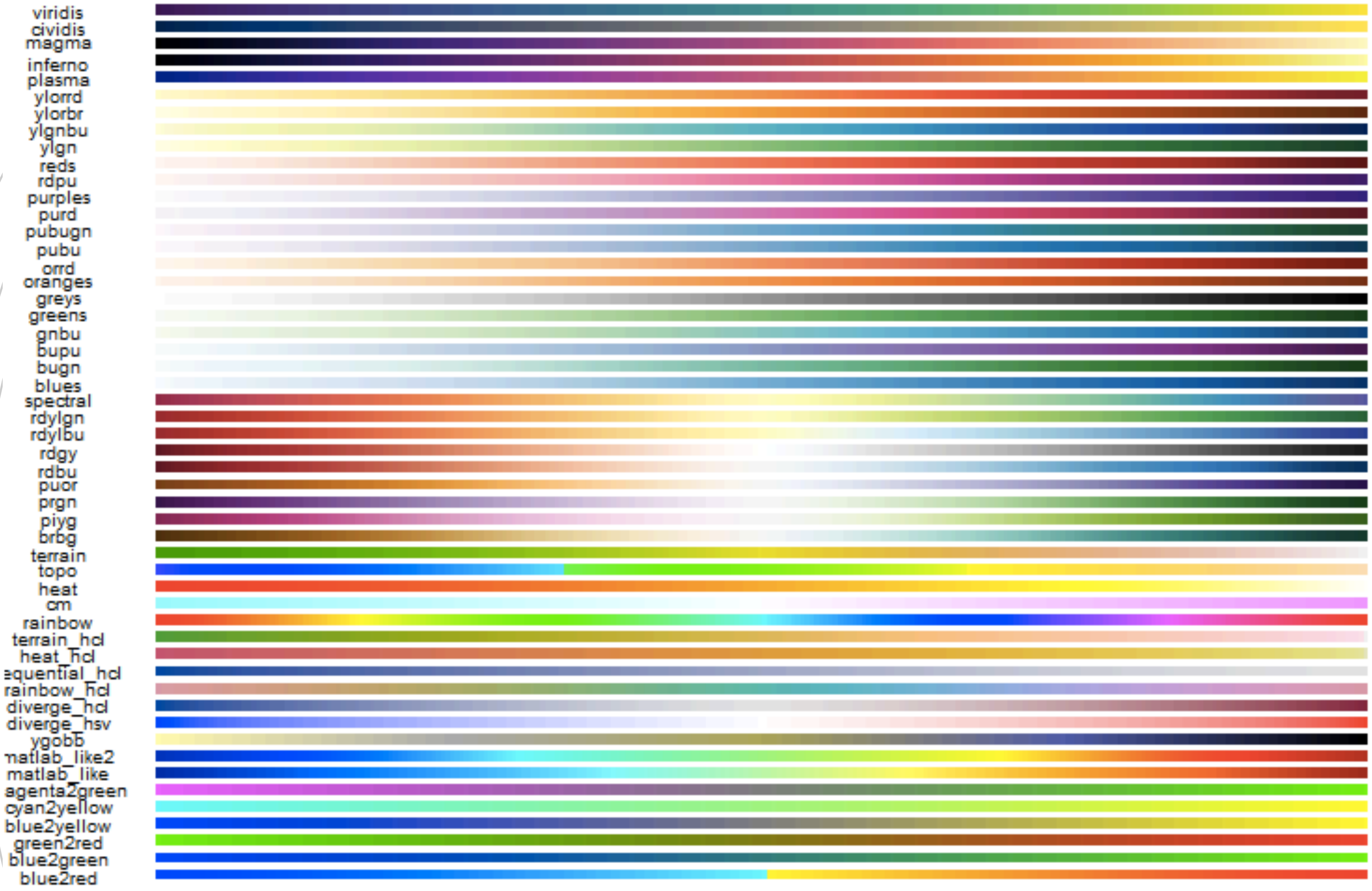
To look at multiple secondary features of interest:

To use:

interact_multi(count_data, main_info_data, sec_info_data, dim_red_meth = c("UMAP", "PCA"), your_palette, main_info_name, sec_info_name)

Interact_multi





Colour Palette Options:

Exercice!

1. Using the interact_mapper package, try to generate a dimension reduction plot of your choice using the read count data provided on the patients and their clinical features to see if any clinical features relate to the structure of the data
2. Using the interact_mapper package, try to generate a dimension reduction plot of your choice using the read count data provided on the patients to see if your gene of interest relates to the structure of the data

Questions to ask yourself when you're designing figures:

What do I look for in figures…

In a paper?

In a presentation?

Who am I making this figure for/what is my audience?
What are their needs?

Why am I making this figure? What question am I answering in making it?
What is the key message here?

Is my message clear and understandable? Is my data easily seen and interpretable?

Is it visually appealing?

Cleaning up your plot

how to display several plots together

Before making your plots, type in:

```
par(mfrow=c(a,b))  #where a is the number of rows
and b is the number of columns in your grid of
plots
```

Then make your plots!

To reset your display, type in:

```
dev.off()
```