# Data viz for scientists

## How a picture tells a 1000s words

21st october 2021

IPOP-UP bioinformatics meetings

Alix Silvert (They/them)

# Graphs are cool !

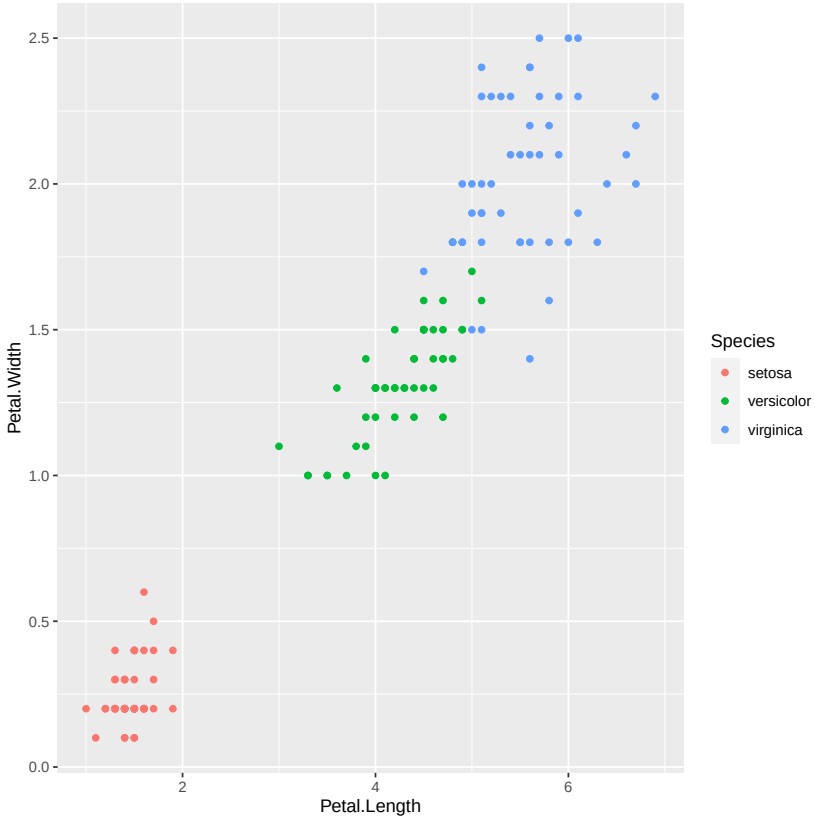
- Eyeball the data
- Summarize data or results
- Reinforce a point
- Beautiful and attention grabbing

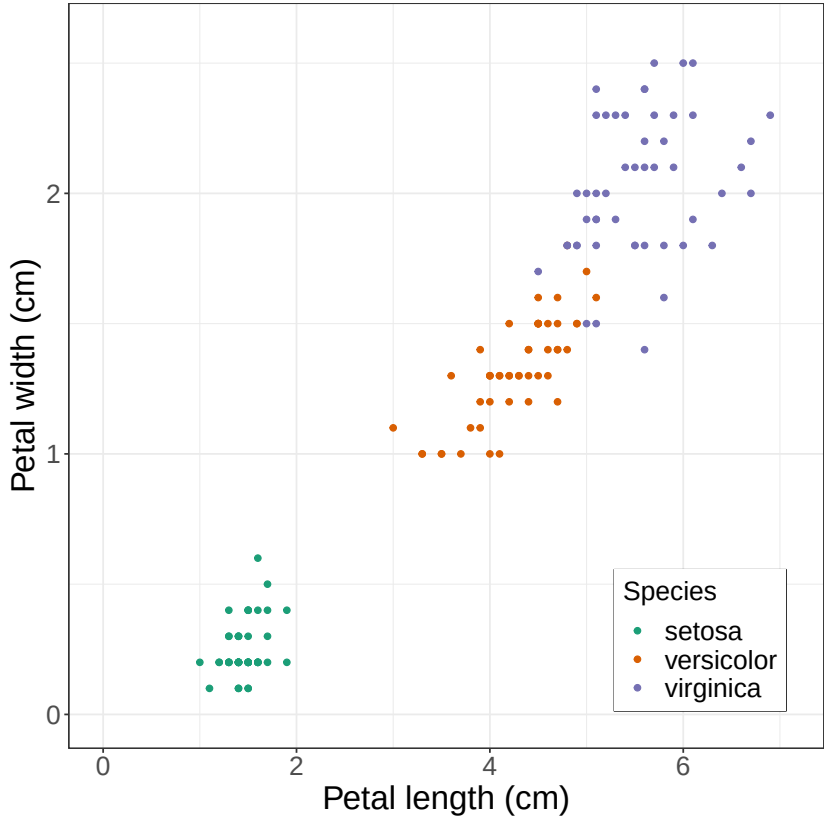
# What am I using ?

- R and ggplot2
  - Easy to use, hard to mess up too bad
- Datasets included in ggplot2 (when I can)
  - Iris
  - Diamonds

# Readability

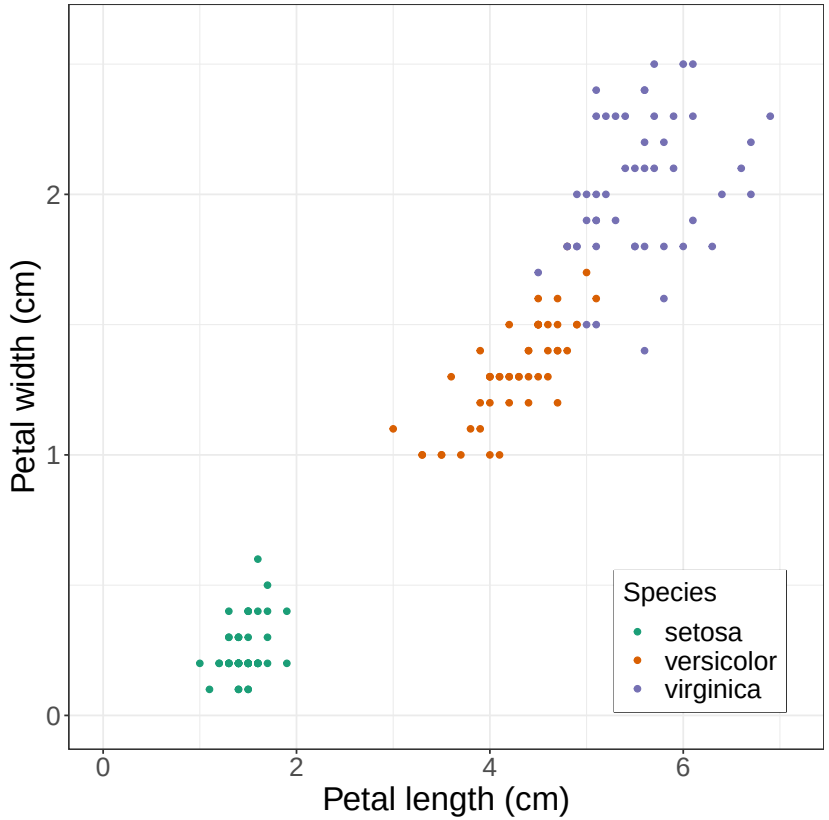
# Basic quality



# What we're aiming at

- Labeled axis
- Readable text
- Efficient use of space (legend)
- Visible colors
- Not the "default" look

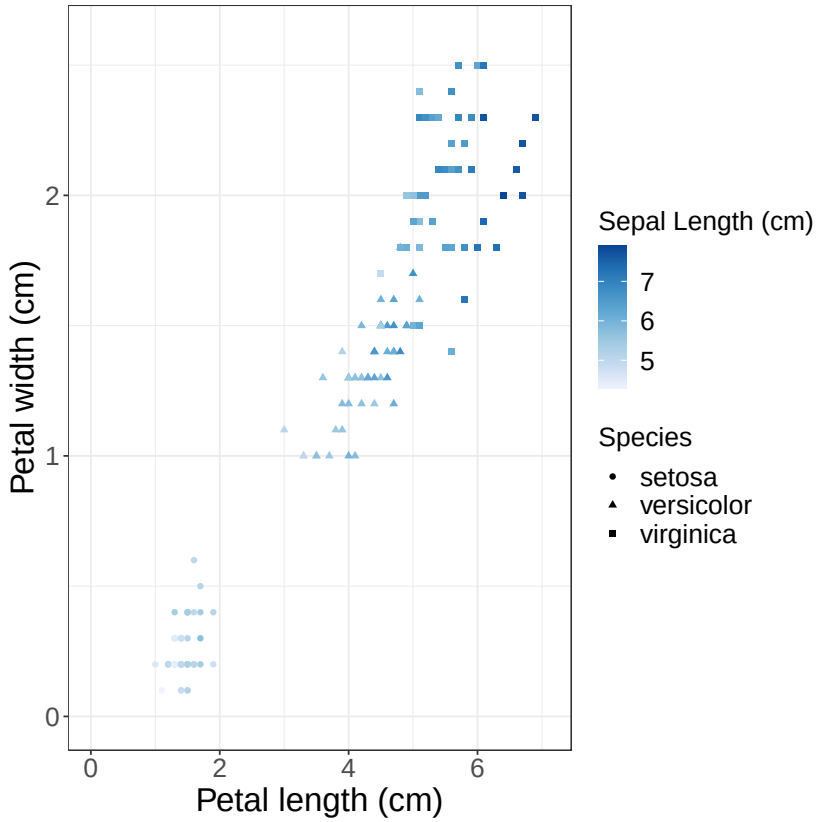


# It takes some work

```r
p <- ggplot(iris, aes(x = Petal.Length, y = Petal.Width, color = Species))
p <- p + geom_point()
```

```r
q <- ggplot(iris, aes(x = Petal.Length, y = Petal.Width, color = Species))
q <- q + geom_point()
q <- q + theme_bw()
q <- q + xlim(c(0,7.1)) + ylim(c(0, 2.6))
q <- q + xlab("Petal length (cm)") + ylab("Petal width (cm)")
q <- q + scale_color_brewer(name = "Species", palette = "Dark2")
q <- q + theme(
  axis.text = element_text(size = 16),
  axis.title = element_text(size = 20),
  legend.text = element_text(size = 16),
  legend.title = element_text(size = 16),
  legend.justification = c(1, 0),
  legend.position = c(0.95, 0.05),
  legend.box.background = element_rect(colour = "black")
)
```

# Keep it simple !




- One graph, one idea
- If you need more, make several graphs
- NO 3D !
- (Dimensionality reduction is it’s own subject)

# Where will this figure be ?

- How are the colors going to show ?
- Is the size adapted ?
- Will the text be readable ?
- What format of output is best ?

# Image formats

## Vectorial

- Bunch of formulas
- No compression
- Friendly to post-R improvements

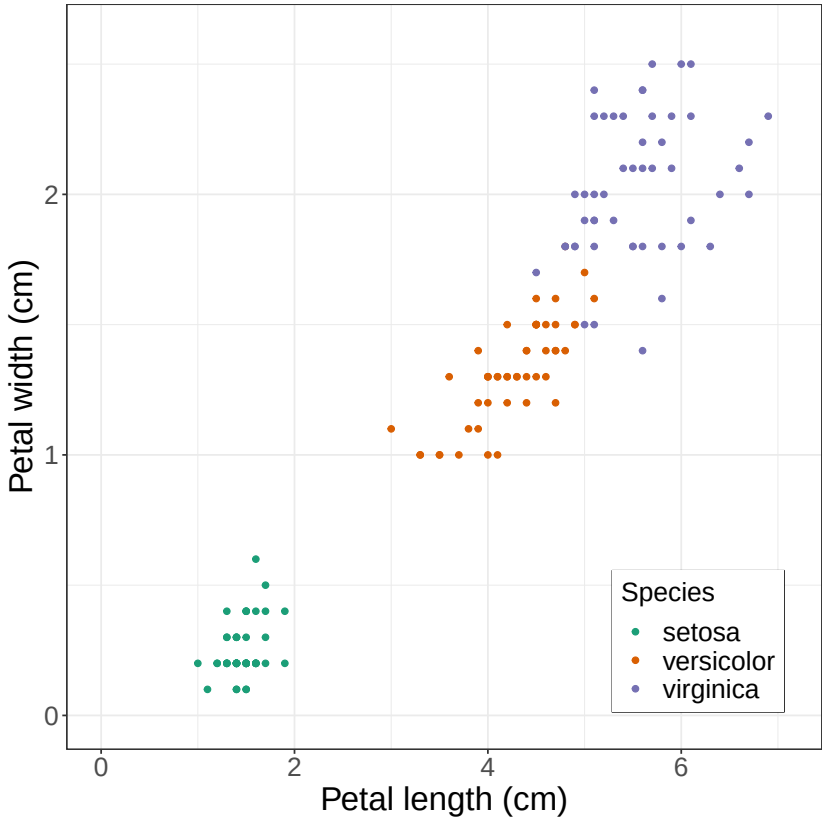
## Raster

- Bunch of pixels
- Compression algorithm

# Vectorial modification

.svg format / inkscape

- Select every dot of a single color
- Modify your text
- Resize without lost



# Readability in short

- Labeled axis and legend
- Clear color differences
- Appropriate font and font size
- One idea per graph
- Plot your graph for your output
- Use vectorial format when possible

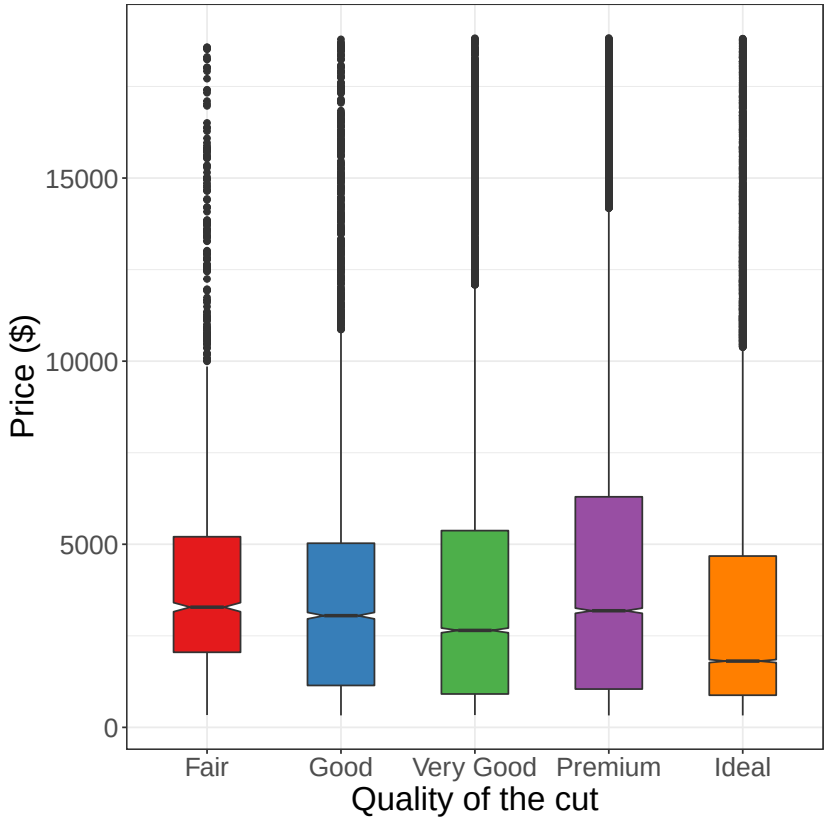
# The Dark Side of dataviz

# Dark side ?

- Graph is clear, but misleading
- Exploiting a bias in thinking
- Playing on a hidden assumption

it is not always done on purpose

# An innocent example



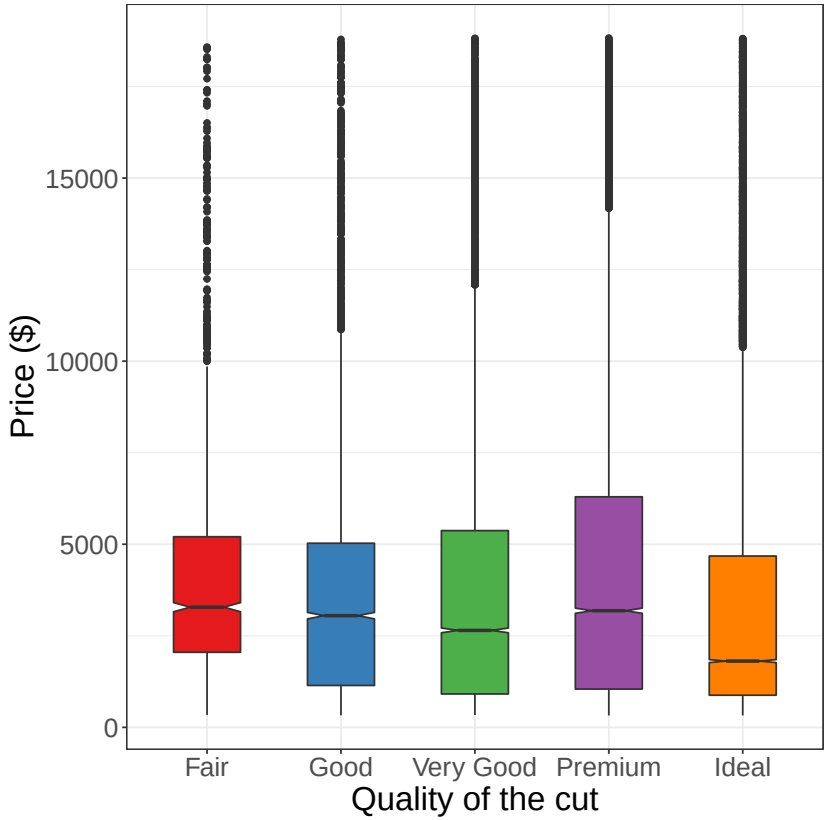

What do the whiskers represent ?

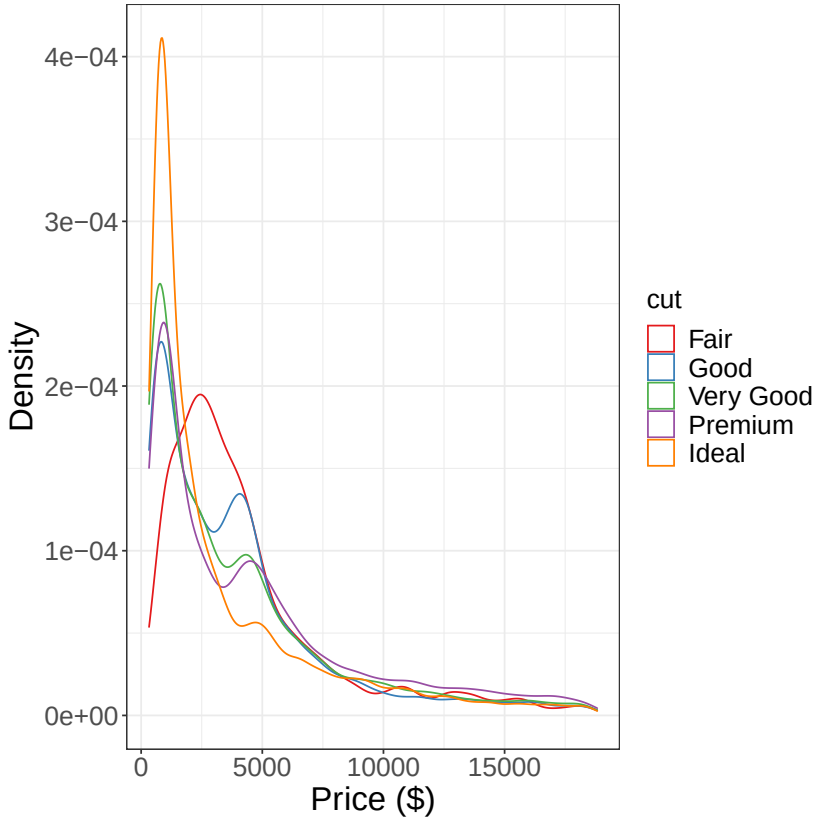
# An innocent example

```
The upper whisker extends from the hinge to the largest value no
further than 1.5 * IQR from the hinge (where IQR is the
inter-quartile range, or distance between the first and third
quartiles). The lower whisker extends from the hinge to the
smallest value at most 1.5 * IQR of the hinge. Data beyond the end
of the whiskers are called "outlying" points and are plotted
individually.
```

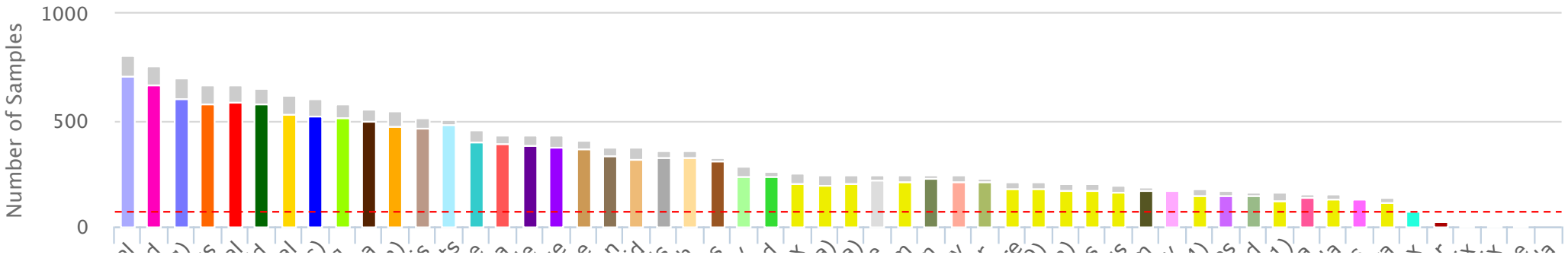
Did you know it ? Would you have checked ?

# Also : wrong type of graph



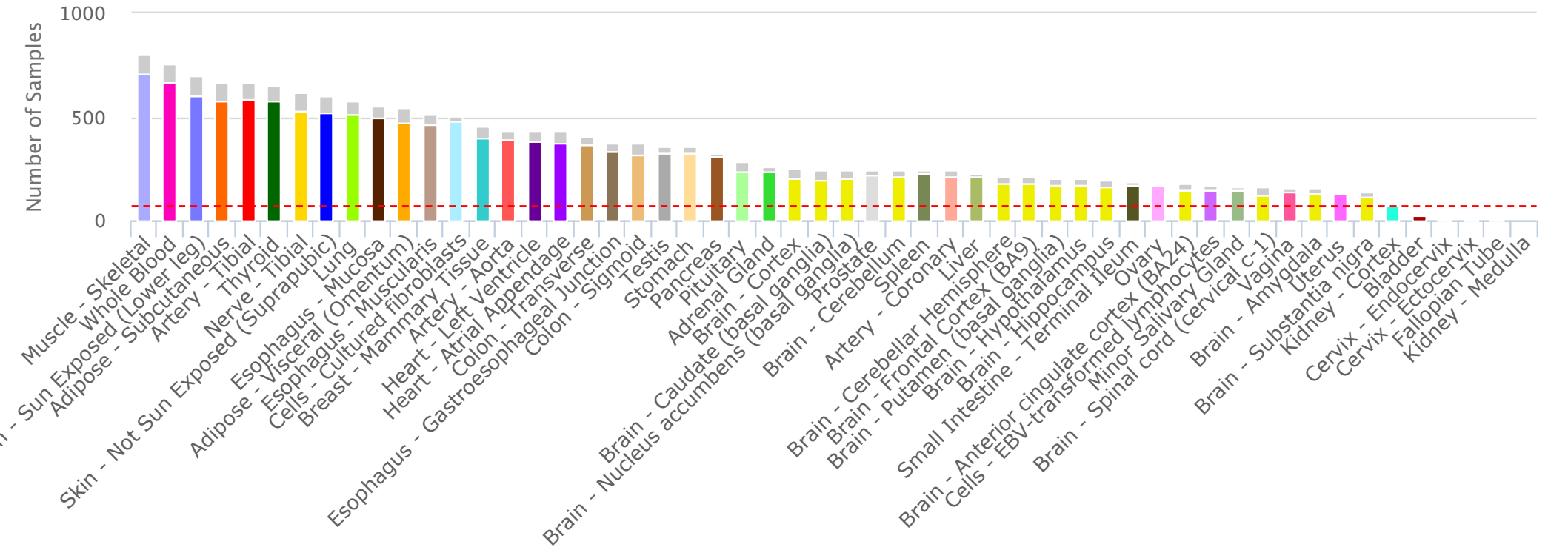
# Colors have meaning (GTEx)



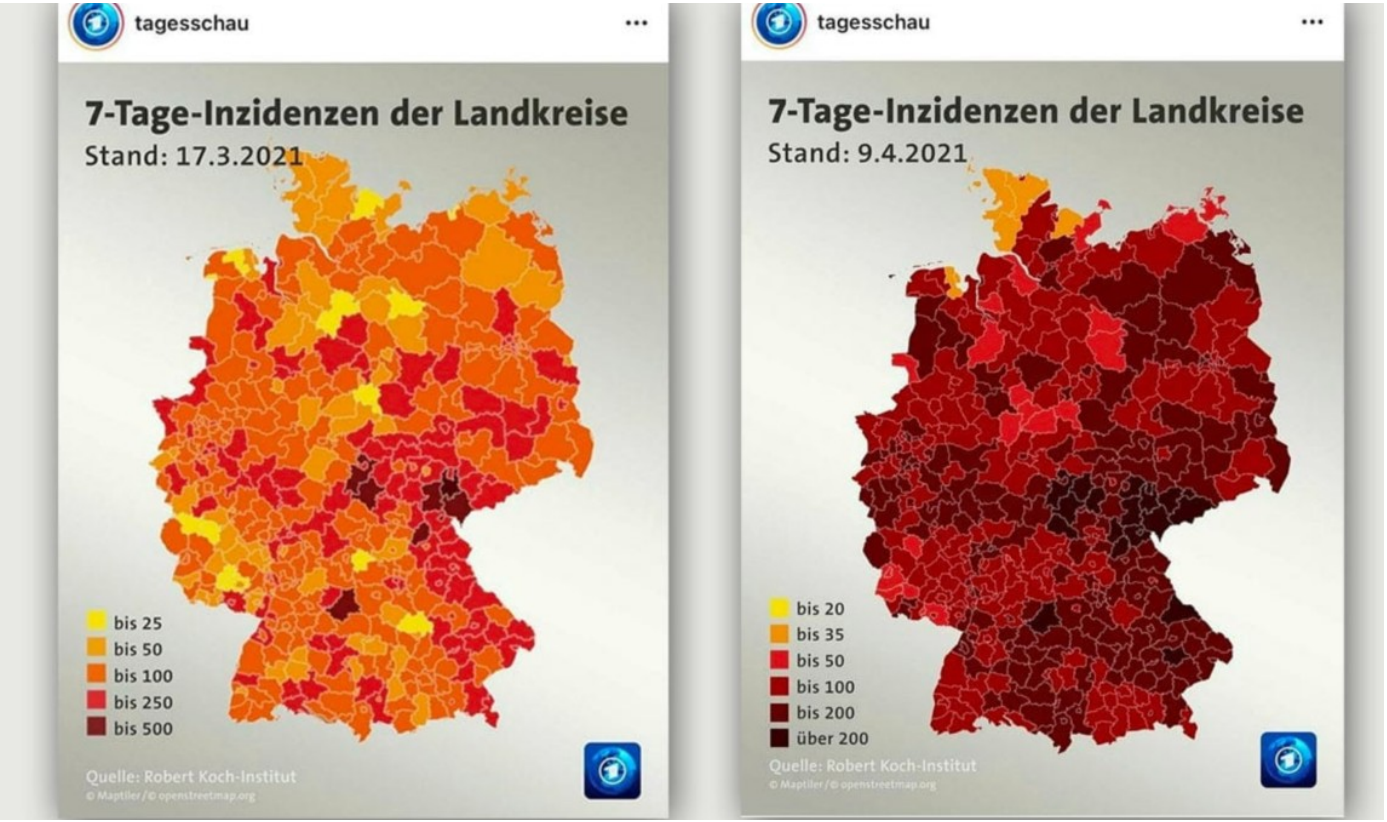

What tissue are related ?

Which one is “Whole Blood”

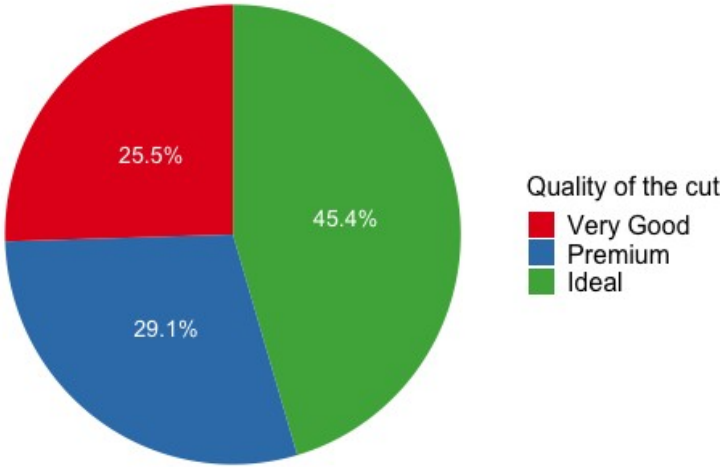
# Colors have meaning (GTEx)

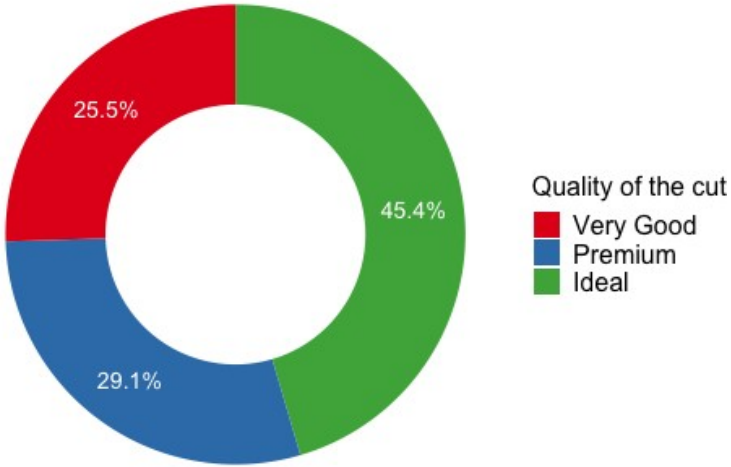


# Distorted/misleading color scale



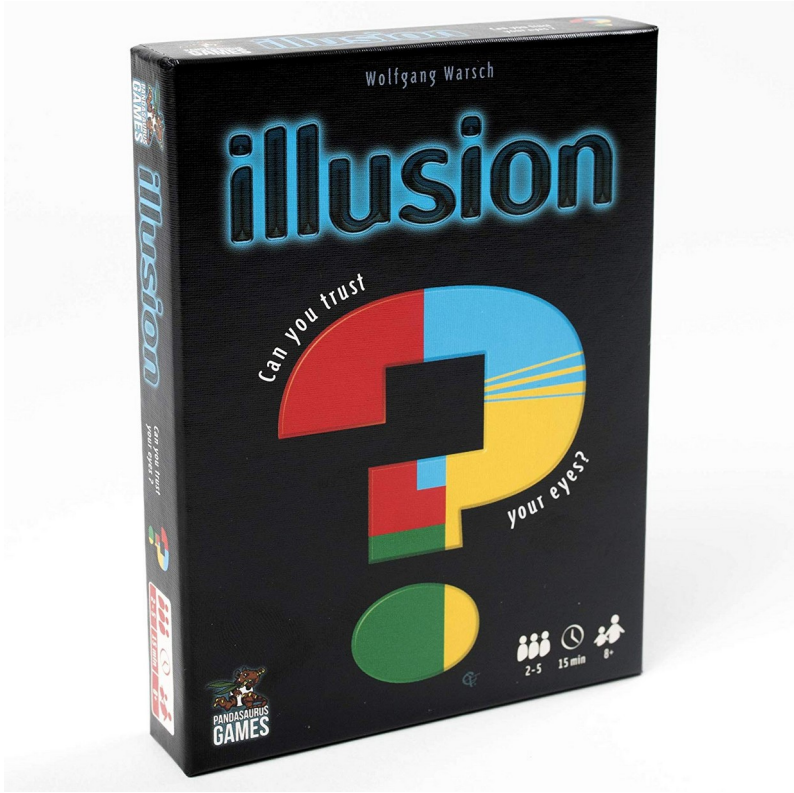
# Areas and angles are hard to read



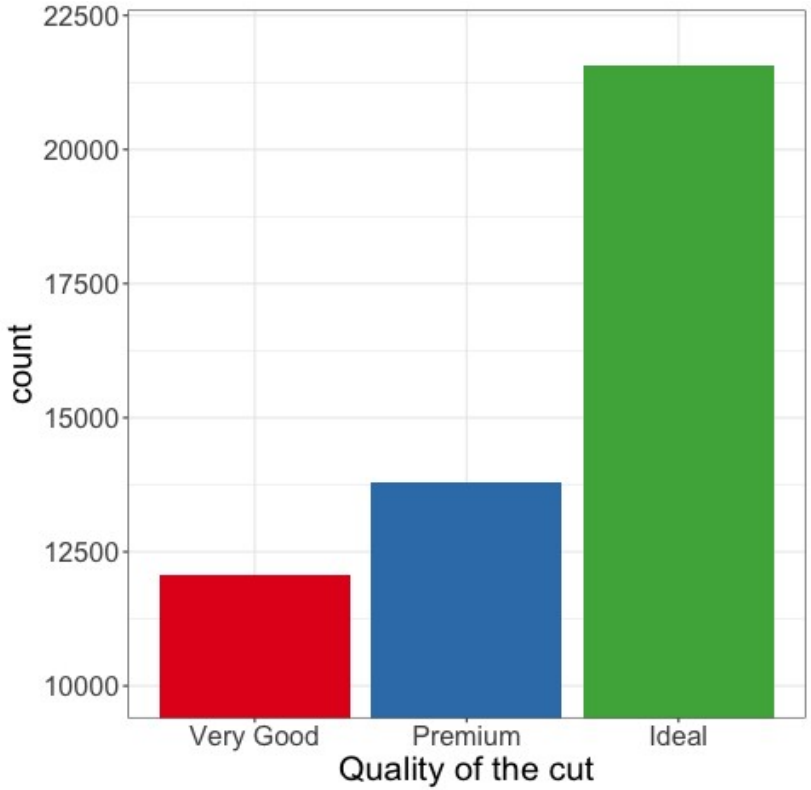



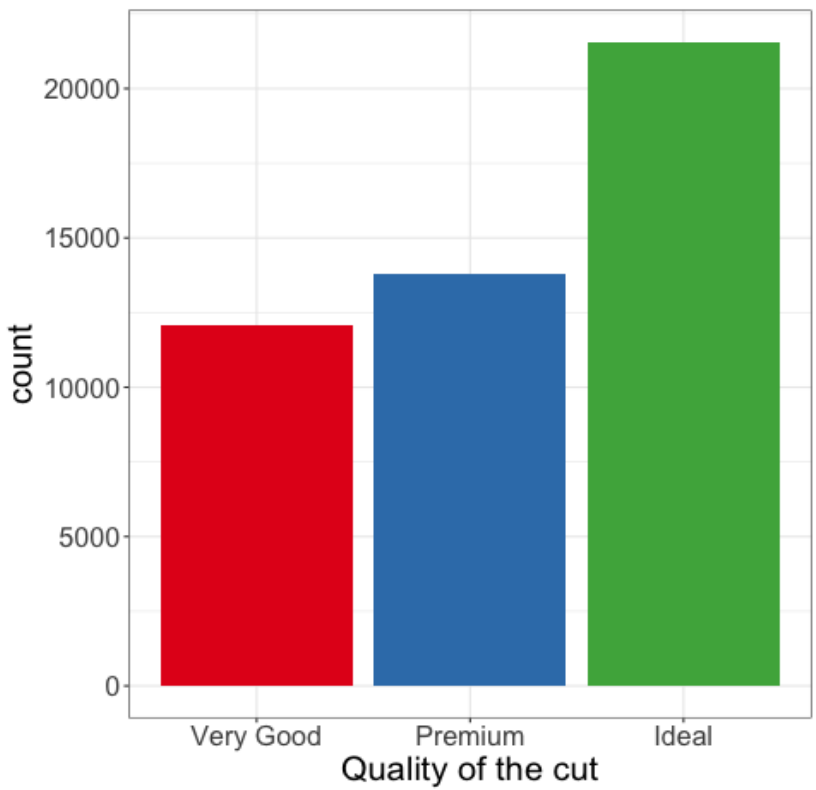
# Areas and angles are hard to read

There’s literally a game about that

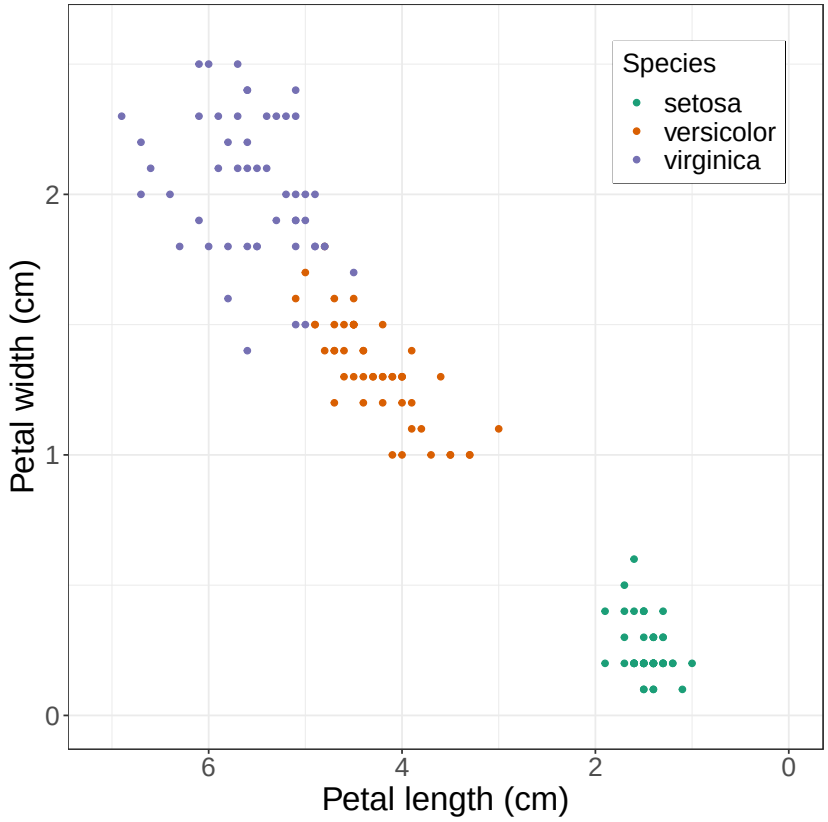


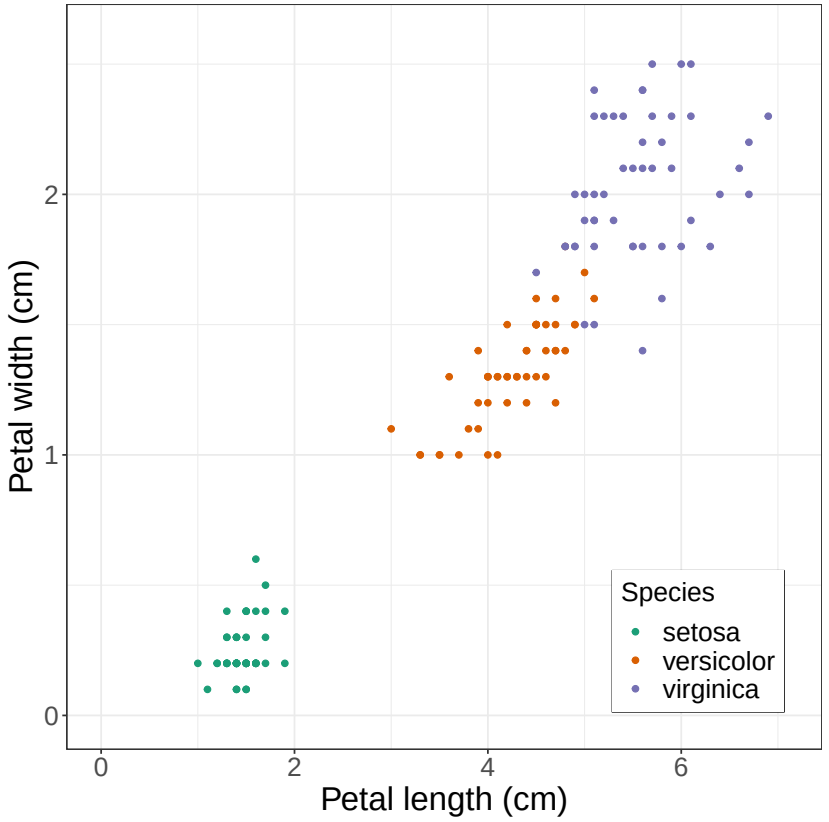
# Truncated axis



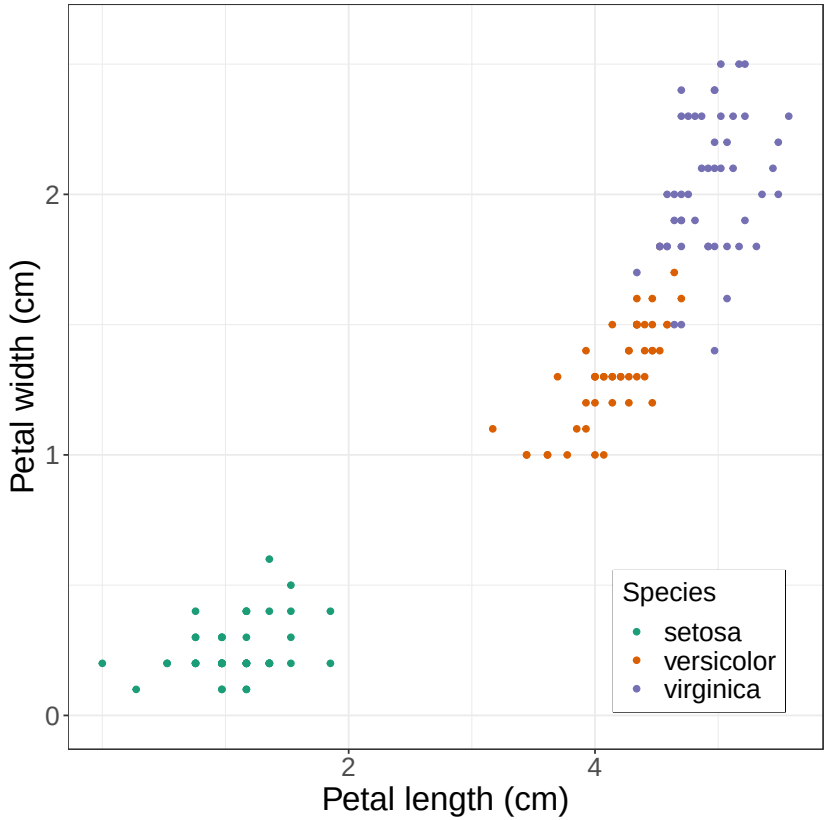



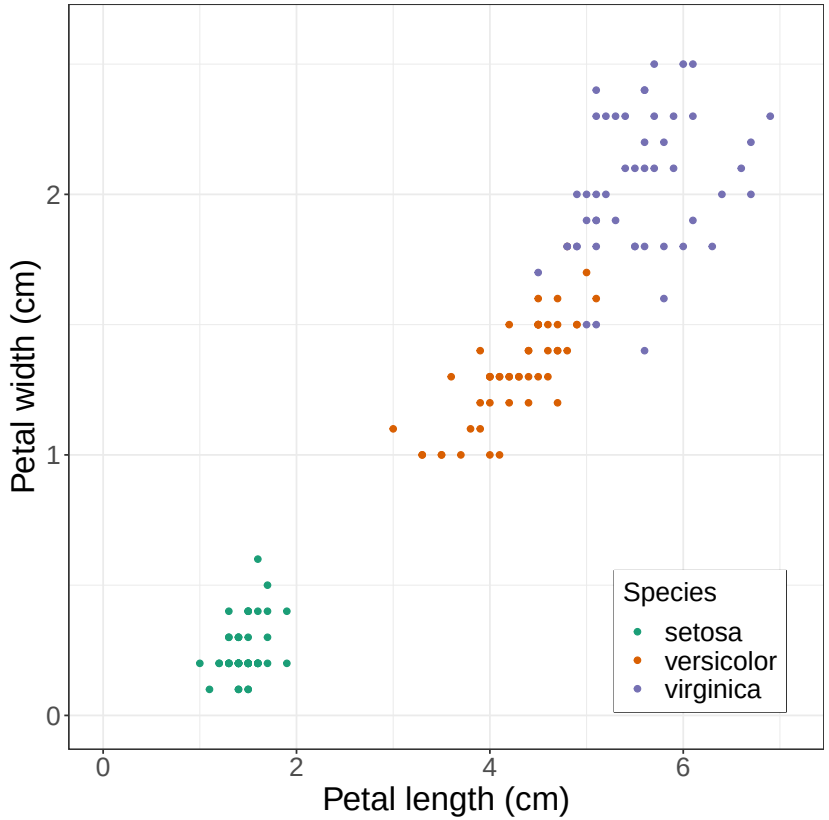
# Inverted axis



# Non-linear axis (left one is log2)



# Inverted axis, or distorted axis

- Avoid it as much as possible
- Plot the modified variable when needed

  (i.e. log2(RPKM + 1))
- Choose your variable so that what goes up and down is intuitive

# In conclusion

- Be sure your graph is understandable
  - Even if people don't listen to you
- Be careful not to mislead the reader
  - It's easier than you think
- https://www.r-graph-gallery.com/

# We are looking for speakers

If you want to communicate on methods or a bioinformatic related subject, contact me :

alix.silvert@u-paris.fr

iPOP-UP