DMI

Data Visualization

Data Mining and Data integration in Biomedicine
Master in Bioinformatics

Janet Piñero
Medbioinformatics Solutions SL
2025-2026

Outline

DMI

Principles of graphical excellence

Visualization with ggplot2

Top 10 worst plots

Venn diagrams

Visualizing networks

References

Hands-on session

Please, update

	A	D
1	Nom	github
2	ITXASO	
5	PABLO	
6	DAVID	
11	DIEGO	
17	REGINA	
19	MARTA	
22	NAHIA	

	A	F	
1	Nom	article title	sessi
2	ITXASO		
3	JÚLIA		
5	PABLO		
6	DAVID		
7	TANNER ALEXANDER		
11	DIEGO		
14	MONTSERRAT		
17	REGINA		
19	MARTA		
22	NAHIA		
26	YANGXIN		
27			

	A	G
1	Nom	session 1
2	ITXASO	
10	LAIA	
14	MONTSERRAT	
19	MARTA	
22	NAHIA	

Link to the file here

When choosing a publication, please check that it is not already chosen by another team

Visualization of Biomedical Data: misconceptions

DMI

Misconception 1: "**The goal of data visualization is to impress**."

We sometimes think of data visualization as purely aesthetic, adding an optional wow factor not present in the data itself. This can be true when creating artwork (e.g., a cover figure), but the role of data visualization in research is almost exactly the opposite: It is a necessary step, aimed at clearly revealing patterns in data.

https://www.annualreviews.org/doi/10.1146/annurev-biodatasci-080917-013424

Visualization of Biomedical Data: misconceptions

DMI

Misconception 2: “**Data visualization is easy.**”

Well-designed visualizations can be so easy to understand and use that we are misled into thinking they must have been easy to create. However, most graphs are simple, but their invention was neither simple nor obvious.

https://www.annualreviews.org/doi/10.1146/annurev-biodatasci-080917-013424

Visualization of Biomedical Data: misconceptions

DMI

Misconception 3: "**Studying data visualization is unnecessary.**"

Underestimating the difficulty of data visualization can lead us to overestimate our current skills and conclude that we would gain little benefit from investing time, effort, or money in training or study.

https://www.annualreviews.org/doi/10.1146/annurev-biodatasci-080917-013424

Principles of Graphical Excellence

DMI

- Graphical excellence is the well-designed presentation of interesting data-a matter of *substance*, of *statistics*, and of *design*
- Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency.
- Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space

Tufte ER. The Visual Display of Quantitative Information.

Principles of Graphical Excellence

DMI

- Show the data
- Induce the viewer to think about the substance of the findings rather that the methodology, the graphical design, or other aspects
- Avoid distorting what the data have to say
- Present many numbers in a small space, i.e., efficiently
- Make large data sets coherent
- Encourage the eye to compare different pieces of data
- Reveal the data at several levels of detail, from a broad overview to the fine structure
- Serve a clear purpose: description, exploration, tabulation, or decoration
- Be closely integrated with the statistical and verbal descriptions of the data set
- *Avoid too many superimposed elements, such as too many curves in the same graphing space.*
- *Find the right aspect ratio and scaling to properly bring out the details of the data.*
- *Avoid having the data all skewed to one side or the other of your graph.*

From E. R. Tufte. The Visual Display of Quantitative Information, 2nd Edition. Graphics Press, Cheshire, Connecticut, 2001.

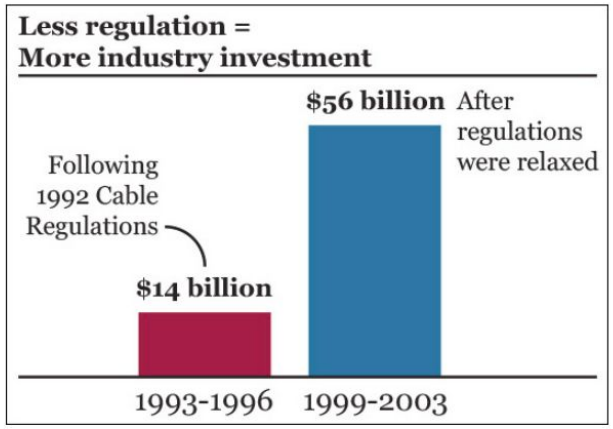
Principles of Graphical Excellence

- Exclude unneeded dimensions
- Omit "chart junk" (term from E.R. Tufte) and unnecessary ink
- Present data in a way to facilitate comparisons
- Make efficient use of space
- Select the best graph type
- Show uncertainty
- Avoid 3-D plots if the third dimension does not add information (Much easier for the eye to compare data/results without the added unnecessary dimension)

The Five Qualities of Great Visualizations

DMI

1. **It is truthful**, as it's based on thorough and honest research.
2. **It is functional**, as it constitutes an accurate depiction of the data, and it's built in a way that lets people do meaningful operations based on it (seeing change in time).
3. **It is beautiful**, in the sense of being attractive, intriguing, and even aesthetically pleasing for its intended audience—scientists, in the first place, but the general public, too.
4. **It is insightful**, as it reveals evidence that we would have a hard time seeing otherwise.
5. **It is enlightening** because if we grasp and accept the evidence it depicts, it will change our minds for the better.



The Truthful Art: Data, Charts, and Maps for Communication, Alberto Cairo

Exploratory vs expository graphs

DMI

Exploratory graphs

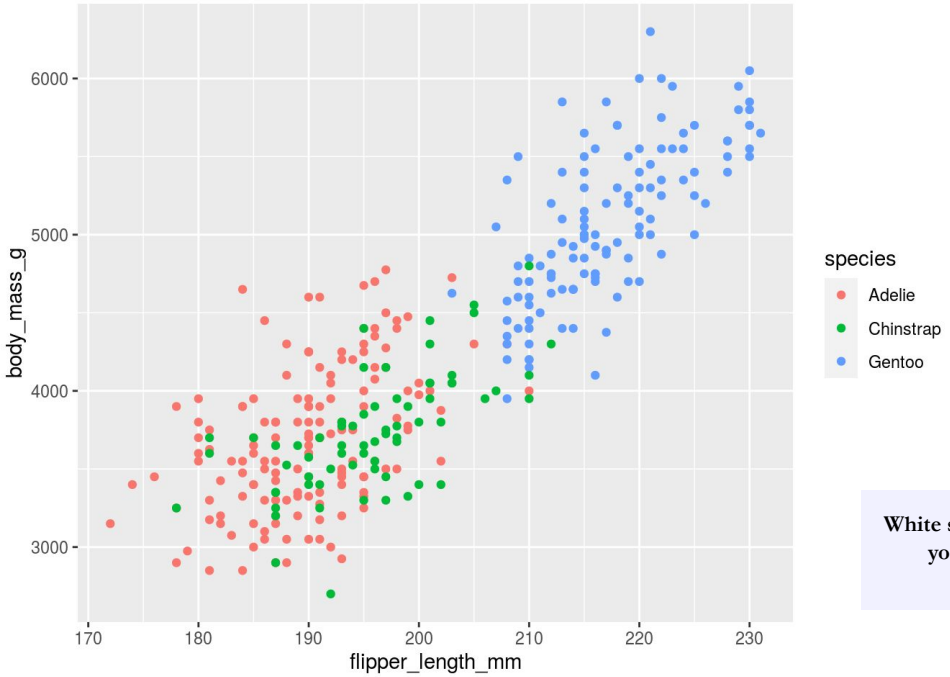
- Are made quickly and you should make a ton of them!
- Used discover structure, patterns, errors in the data for you personally
- The only audience is you so you shouldn't worry about things like fonts, axis labels, titles, legends, and labeling

Expository graphs

- Are shared with an audience.
- Used to communicate findings. Keep in mind that: People spend way less time reading your work than you think
- Ideally they should be self-contained
- Clear, large axis labels; color and size carefully used for communication; minimal abbreviations in axis labels and legends; shows the data; have figure captions with a declarative summary statement and self-sufficient labeling; and highlights take home messages with titles or annotation.

Exploratory vs expository graphs

DMI



White space is like garlic; take the amount you think you need, then triple it.
@W_R_Chase

Exploratory vs expository graphs

DMI

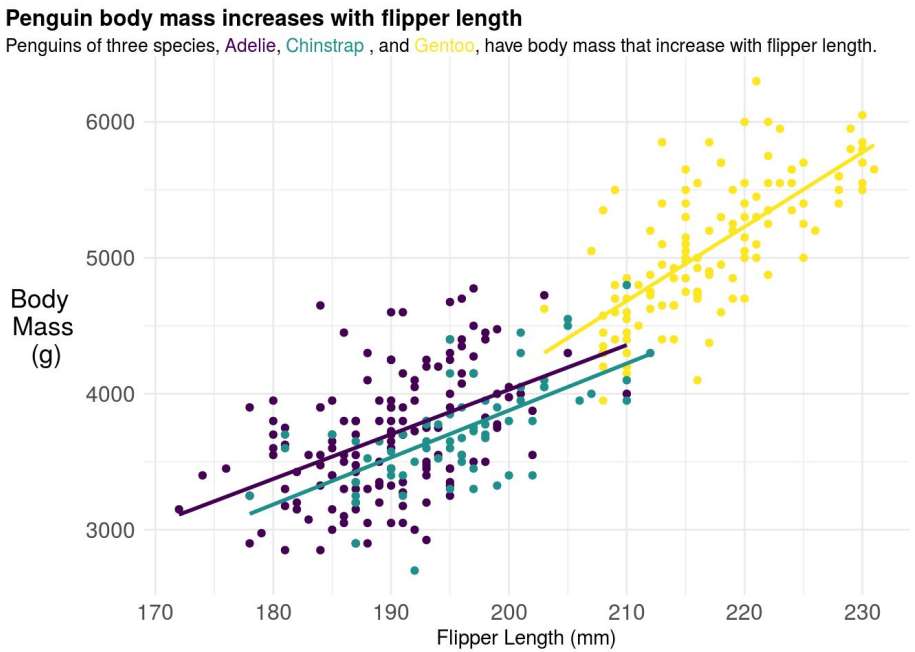


Figure 1. Penguin body mass increases with flipper length across three penguin species A plot of body mass in grams (g) versus flipper length in millimeters (mm) for three Penguin species: Adelie (red), Chinstrap (green), and Gentoo (blue). A separate linear regression fit to each species type is also shown, highlighting that penguin body mass increases with flipper length at different rates across different species.

DMI

The ggplot2 package

Why ggplot2?

DMI

ggplot2 is a data visualization package for R developed by Hadley Wickham that provides a structured approach to graphing.

Pros

- Standardized method for plotting.
- Publication quality plots.
- Allows creation of relatively complex plots with ease.
- Dominant plotting package in R.

Cons

- Might be a little bit difficult to understand at the beginning.
- Sometimes you might get lost between all the things you can change.

ggplot2 package: data

The functions in the ggplot2 package build up a graph in layers.

```
# install.packages("mosaicData")
# load packages
library(ggplot2)
library(mosaicData)
data(CPS85)
# specify data
ggplot(data = CPS85)
```

wage	educ	race	sex	hispanic	south	married	exper	union	age	sector
9.0	10	W	M	NH	NS	Married	27	Not	43	const
5.5	12	W	M	NH	NS	Married	20	Not	38	sales
3.8	12	W	F	NH	NS	Single	4	Not	22	sales
10.5	12	W	F	NH	NS	Married	29	Not	47	clerical
15.0	12	W	M	NH	NS	Married	40	Union	58	const
9.0	16	W	F	NH	NS	Married	27	Not	49	clerical

Data in a tidy (aka long) format.

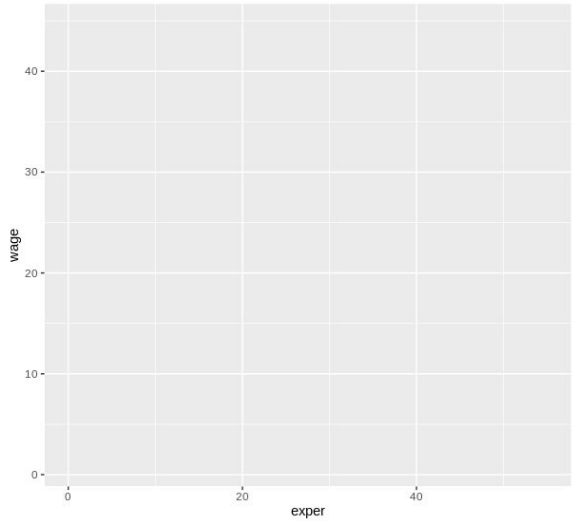
1. Each variable forms a column.
2. Each observation forms a row.

You can convert data from wide to long format using reshape2::melt()

ggplot2 package: aesthetics mappings

The functions in the ggplot2 package build up a graph in layers.

```
# install.packages("mosaicData")
# load packages
library(ggplot2)
library(mosaicData)
data(CPS85)
# specify aesthetics
ggplot(data = CPS85,
       mapping = aes(x = exper, y = wage))
```



The **aes**thetics argument indicates how we map the data columns to visual elements or parameters.

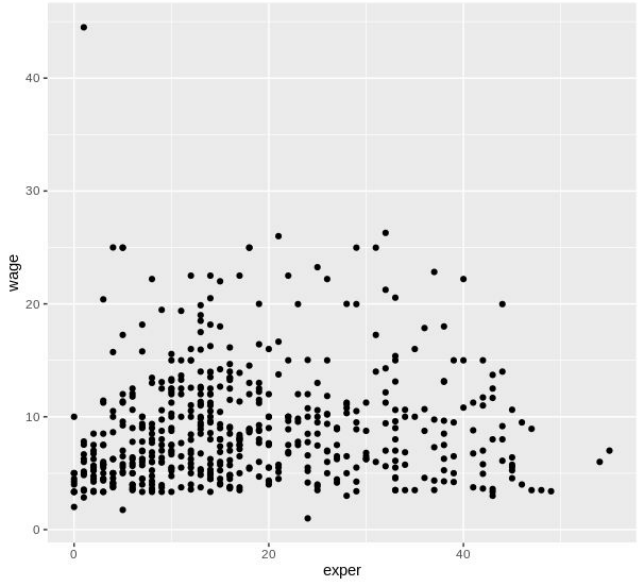
- Variable 1 → x
- Variable 2 → y
- Variable 3 → color, fill, shape, etc.

ggplot2 package: geometries

The functions in the ggplot2 package build up a graph in layers.

```
# install.packages("mosaicData")
# load packages
library(ggplot2)
library(mosaicData)
data(CPS85)
# specify aesthetics
ggplot(data = CPS85,
       mapping = aes(x = exper, y = wage))

# specify geometries
ggplot(data = CPS85,
       mapping = aes(x = exper, y = wage)) +
  geom_point()
```

The **geom**etries indicate which geometric shapes should be plotted by using the parameters and variables defined in aes().



ggplot2 package: geometries

DMI

The geometries indicate which geometric shapes should be plotted by using the parameters and variables defined in aes().

These are just some of all the available geom_

One variable:
Discrete:
geom_bar()

Continuous
geom_histogram()

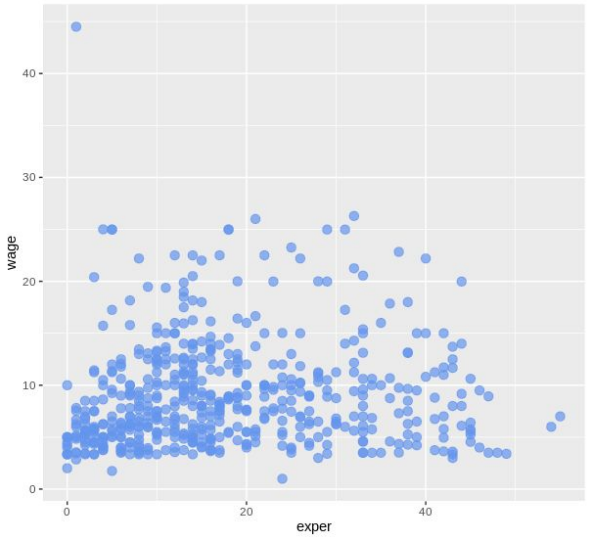
Two variables:
Both continuous:
geom_point().

One continuous, one discrete:
geom_boxplot().

ggplot2 package: geometries

The functions in the ggplot2 package build up a graph in layers.

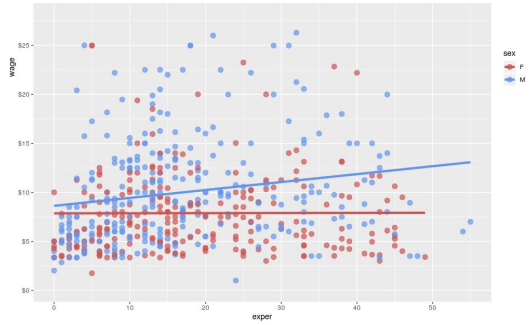
```
# install.packages("mosaicData")
# load packages
library(ggplot2)
library(mosaicData)
data(CPS85)
# specify aesthetics
ggplot(data = CPS85,
       mapping = aes(x = exper, y = wage))

# specify geometries
ggplot(data = CPS85,
       mapping = aes(x = exper, y = wage)) +
  geom_point()

# make points blue, larger, and semi-transparent:
# color, size, alpha.
ggplot(data = CPS85,
       mapping = aes(x = exper, y = wage)) +
geom_point(color = "cornflowerblue", alpha = .7, size
= 3)
```



ggplot2 package: groupings

DMI

```
# modify the x and y axes and specify the colors to be used
ggplot(data = CPS85,
       mapping = aes(x = exper, y = wage, color = sex)) +
  geom_point(alpha = .7, size = 3) +
  geom_smooth(method = "lm",se = FALSE, size = 1.5) +
  scale_x_continuous(breaks = seq(0, 60, 10)) +
  scale_y_continuous(breaks = seq(0, 30, 5),
                     label = scales::dollar) +
  scale_color_manual(values = c("indianred3","cornflowerblue"))
```



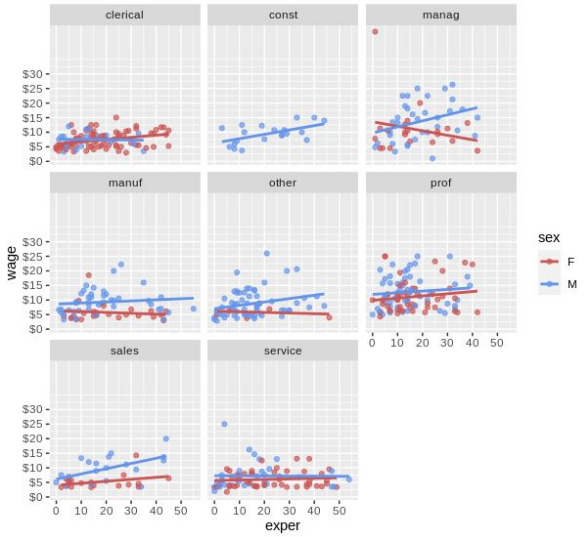
The color = sex option is placed in the aes function, because we are mapping a variable to an aesthetic.

The geom_smooth option (se = FALSE) was added to suppresses the confidence intervals.

ggplot2 package: facets & scales

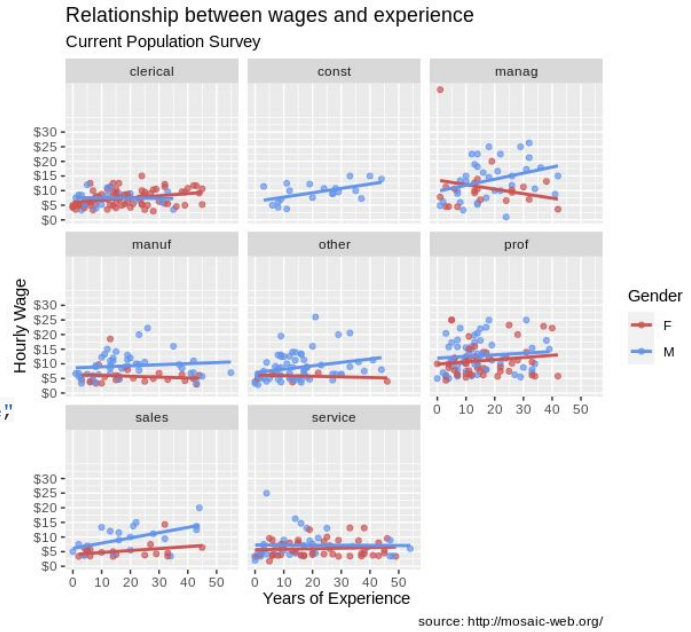
```
# reproduce plot for each level of job sector: facets
ggplot(data = CPS85,
       mapping = aes(x = exper, y = wage, color = sex)) +
  geom_point(alpha = .7) +
  geom_smooth(method = "lm",se = FALSE) +
  scale_x_continuous(breaks = seq(0, 60, 10)) +
  scale_y_continuous(breaks = seq(0, 30, 5),
                     label = scales::dollar) +
  scale_color_manual(values = c("indianred3", "cornflowerblue")) +
    facet_wrap(~sector)
```

- scale_x_discrete() allows you to modify your discrete **x axis**.
- scale_y_continuous() allows you to modify your continuous **y axis**.
- scale_color_manual() if you want to manually adjust the **colors** of your discrete variable.
- scale_fill_gradient() allows you to adjust the colors of your **fill** gradient (for a continuous variable).



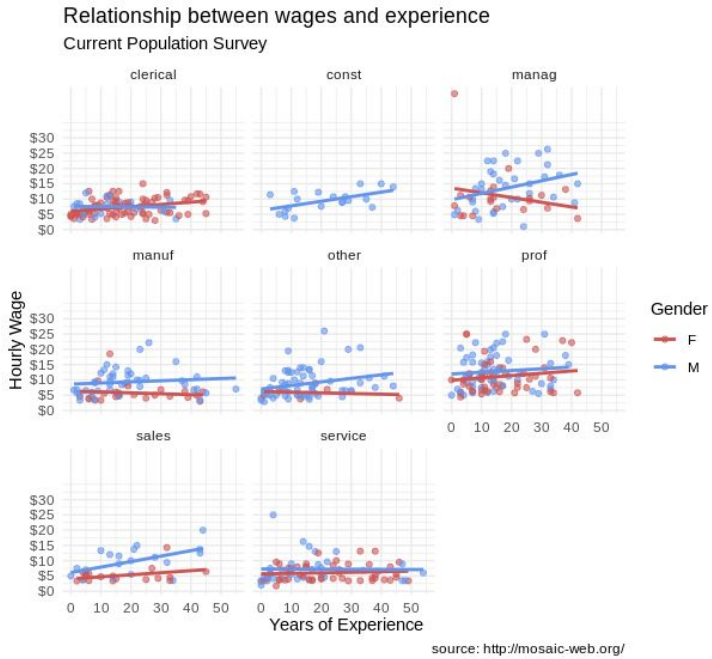
ggplot2 package: labels

```
# add informative labels
ggplot(data = CPS85,
       mapping = aes(x = exper,
                     y = wage,
                     color = sex)) +
  geom_point(alpha = .7) +
  geom_smooth(method = "lm",
              se = FALSE) +
  scale_x_continuous(breaks = seq(0, 60, 10)) +
  scale_y_continuous(breaks = seq(0, 30, 5),
                     label = scales::dollar) +
  scale_color_manual(values = c("indianred3",
                                "cornflowerblue")) +
  facet_wrap(~sector) +
  labs(title = "Relationship between wages and experience",
       subtitle = "Current Population Survey",
       caption = "source: http://mosaic-web.org/",
       x = " Years of Experience",
       y = "Hourly Wage",
         color = "Gender")
```



ggplot2 package: themes

```
ggplot(data = CPS85,
       mapping = aes(x = exper,
                     y = wage,
                     color = sex)) +
  geom_point(alpha = .6) +
  geom_smooth(method = "lm",
              se = FALSE) +
  scale_x_continuous(breaks = seq(0, 60, 10)) +
  scale_y_continuous(breaks = seq(0, 30, 5),
                     label = scales::dollar) +
  scale_color_manual(values = c("indianred3",
                                "cornflowerblue")) +
  facet_wrap(~sector) +
  labs(title = "Relationship between wages and
experience",
       subtitle = "Current Population Survey",
       caption = "source: http://mosaic-web.org/",
       x = " Years of Experience",
       y = "Hourly Wage",
       color = "Gender") +
    theme_minimal()
```



ggplot2 package

Saving your ggplot2 graphs

The function needed for saving ggplot graphs is ggsave().

By default, it will save the last plot you generated in the specified file filename. It will automatically extract the correct format from your filename suffix ("png", "pdf", "svg" among others).

Check all options typing ?ggsave

```
ggsave(filename="plot_example.png")
```

ggplot2 package

DMI

To investigate on your own:

With ggplotly() by Plotly, you can convert your ggplot2 figures into interactive ones.

Example:

```
p <-ggplot(iris, aes(Sepal.Length, Sepal.Width)) +  geom_point(aes(color = Species))
ggplotly(p)
```

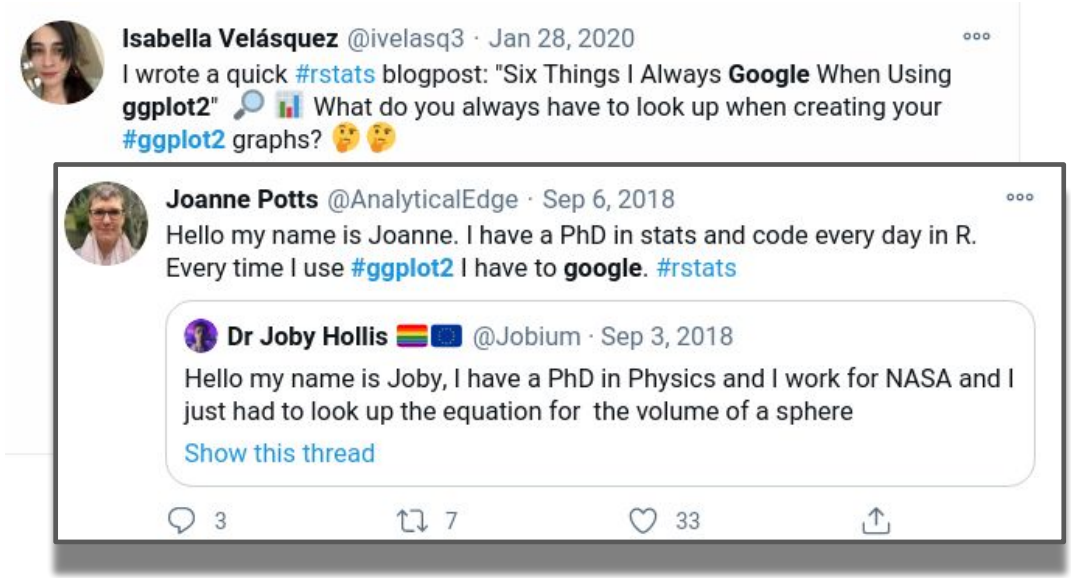
See more at https://plot.ly/ggplot2/

The Evolution of a ggplot

ggplot2 package

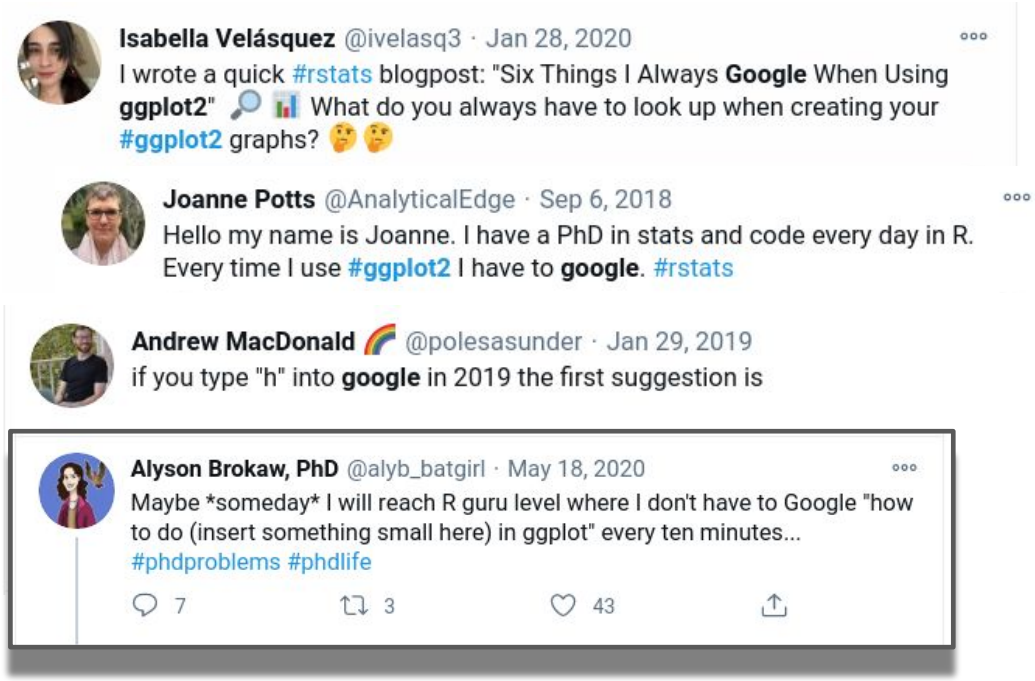
https://ivelasq.rbind.io/blog/things-i-google/

ggplot2 package



ggplot2 package

ggplot2 package



Why do we use color?

Color in data visualization

DMI

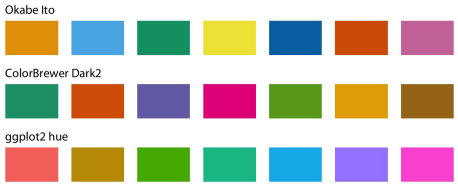
(i) To distinguish groups of data from each other

(ii) To represent data values

(iii) To highlight.

Color scales: distinguishing groups

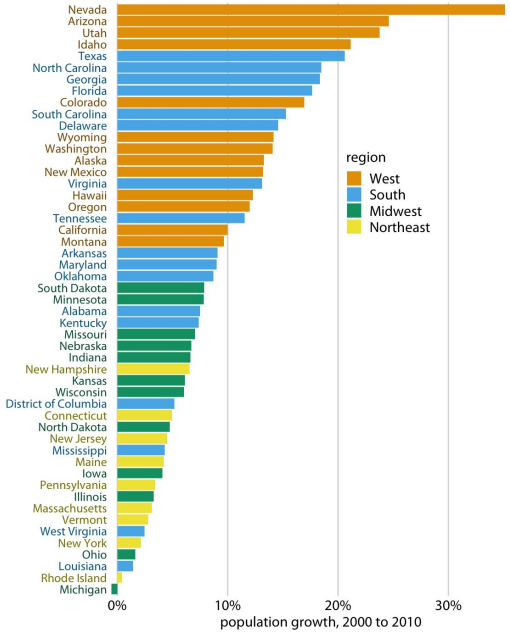
Example qualitative color scales



The Okabe Ito scale is useful for color blind people (https://jfly.uni-koeln.de/color/).

The ColorBrewer Dark2 scale is provided by the ColorBrewer project (http://colorbrewer2.org).

The ggplot2 hue scale is the default qualitative scale in ggplot2.

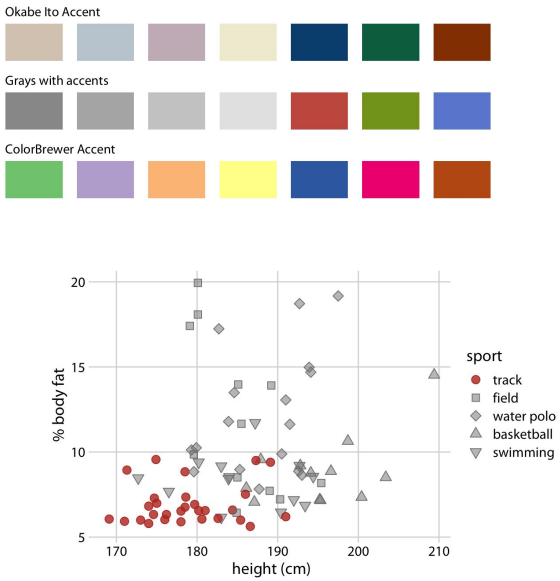


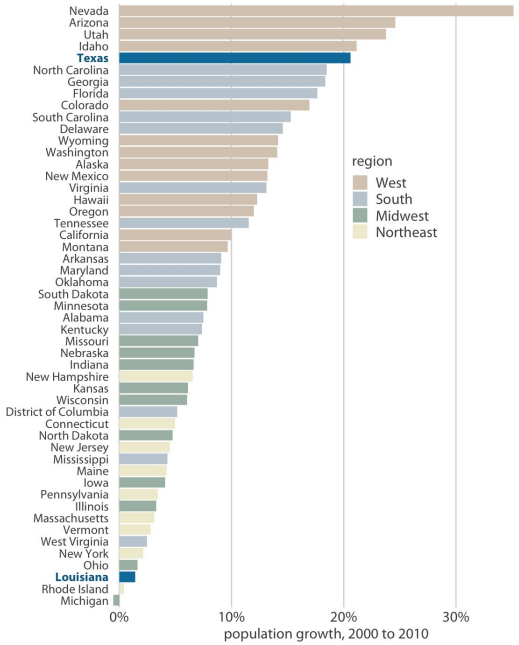
Color scales: representing data values

Sequential scales based on a single hue (e.g., from dark blue to light blue) or on multiple hues (e.g., from dark red to light yellow).
The Heat and Viridis scales are multi-hue scales that vary from dark red to light yellow and from dark blue via green to light yellow, respectively.

diverging color scale: data values deviate in one of two directions relative to a neutral midpoint.
Diverging scale are two sequential scales stiched together at a common midpoint, which usually is represented by a light color
One straightforward example is a dataset containing both positive and negative numbers.

Color scales: highlighting





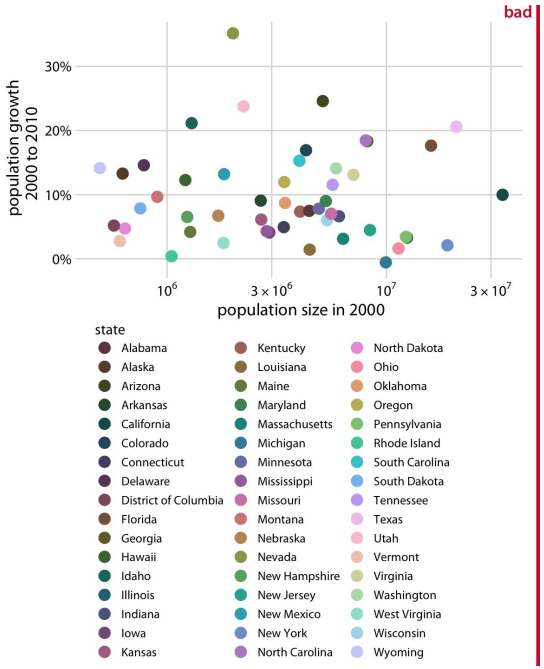
Principles of Graphical Excellence: color

DMI

- Avoid use of multiple colors to "pretty up" a plot; use informative colors and symbol
- Use of too many colors may be distracting
- Use different colors to represent different groupings/categories, *but use them consistently throughout*
- Instead of colors, think about using gray scale, different line styles, or different symbols if the plot will likely be printed in black and white or photocopied

Adapted from Frank E. Harrell Jr. on graphics:
http://biostat.mc.vanderbilt.edu/twiki/pub/Main/StatGraphCourse/graphscourse.pdf

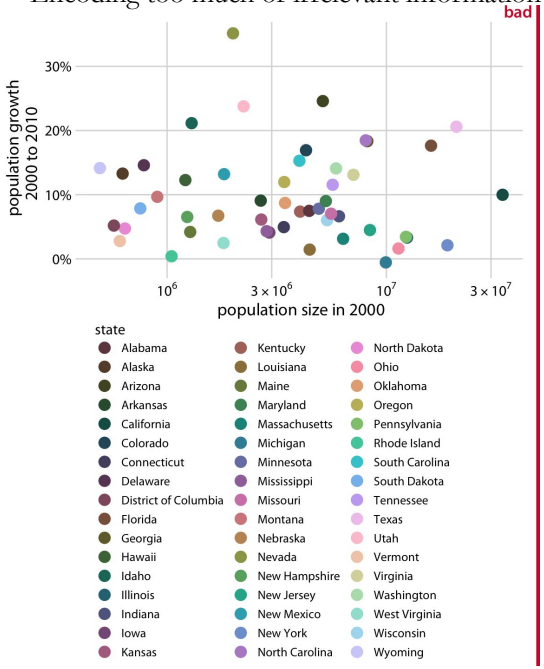
Common pitfalls of color use

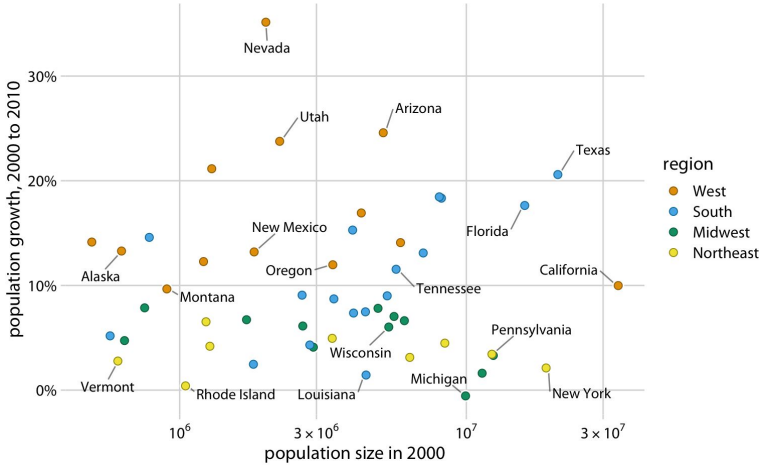


Common pitfalls of color use

DMI

Encoding too much or irrelevant information

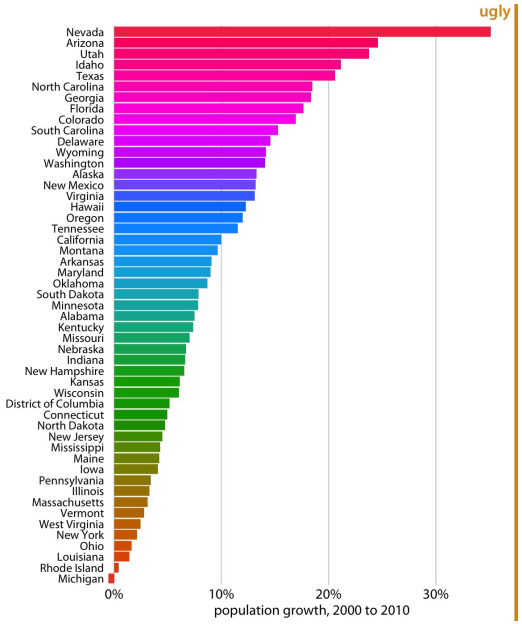




Use direct labeling instead of colors when you need to distinguish between more than about eight categorical items.

Common pitfalls of color use

Coloring for the sake of coloring



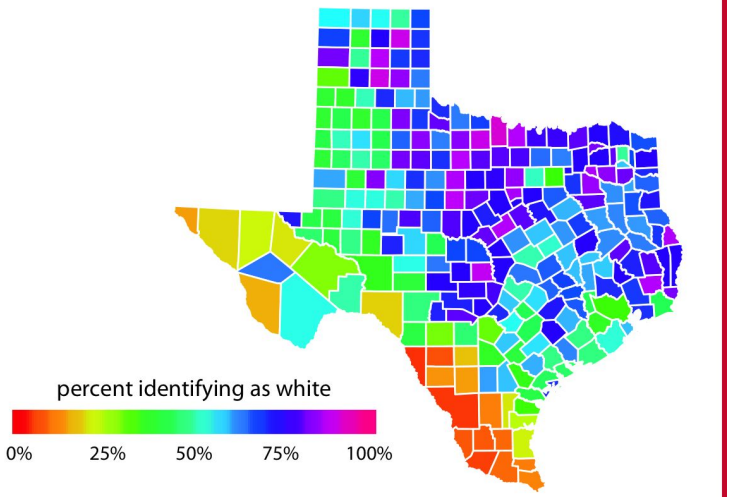
Avoid large filled areas of overly saturated colors. They make it difficult for your reader to carefully inspect your figure.

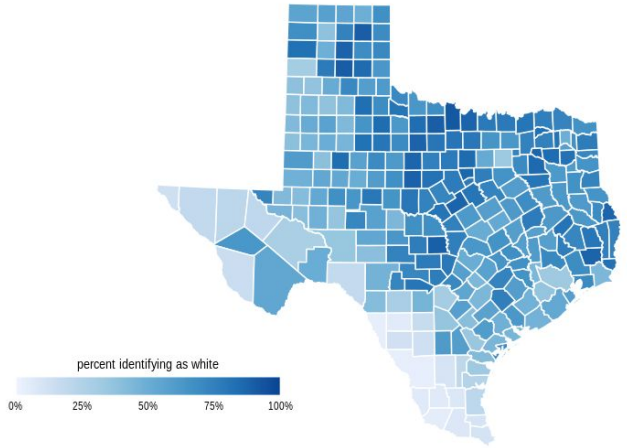
Common pitfalls of color use

DMI

Using non-monotonic color scales to encode data values

bad





Color palettes in R

DMI

a palette in R is simply a vector of colors.

This vector can be coded as hex triplets (Ex. #FFC600) or as R color names.

- **Default palette**

  hist(discoveries, col = palette())

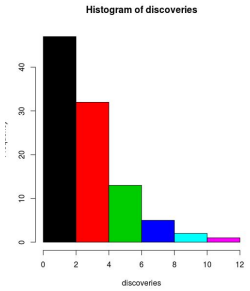
- **package grDevices contains several palettes**

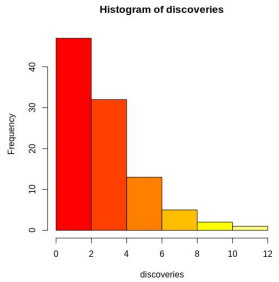
  rainbowcols <- rainbow(6)

  hist(discoveries, col = rainbowcols)

  heatcols <- heat.colors(6)

  hist(discoveries, col = heatcols)





- **colorRampPalette**: Takes a palette of colors and return a function that takes integer arguments and returns a vector of colors interpolating the palette (**grDevices** package)

```
pal <- colorRampPalette(c("red", "yellow"))
pal(10) # "#FF0000" "#FF1C00" "#FF3800" "#FF5500" "#FF7100" "#FF8D00" "#FFAA00"  "#FFC600" "#FFE200" "#FFFF00"
```

Color palettes in R

DMI

- **RColorBrewer palettes: contains 3 types of palettes:**

  Sequential palettes -> ordered data that progress from low to high (gradient).

  The palettes names are : Blues, BuGn, BuPu, GnBu, Greens, Greys,...

  Diverging palettes -> equal emphasis on mid-range critical values and extremes at both ends of the data range.

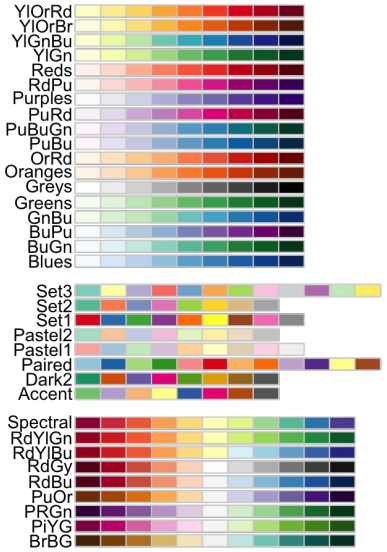
  The diverging palettes are : BrBG, PiYG, PRGn, PuOr, RdBu,...

  Qualitative palettes -> best suited to representing nominal or categorical data. They not imply magnitude differences between groups.

  The palettes names are : Accent, Dark2, Paired, Pastel1…

  How to use it:

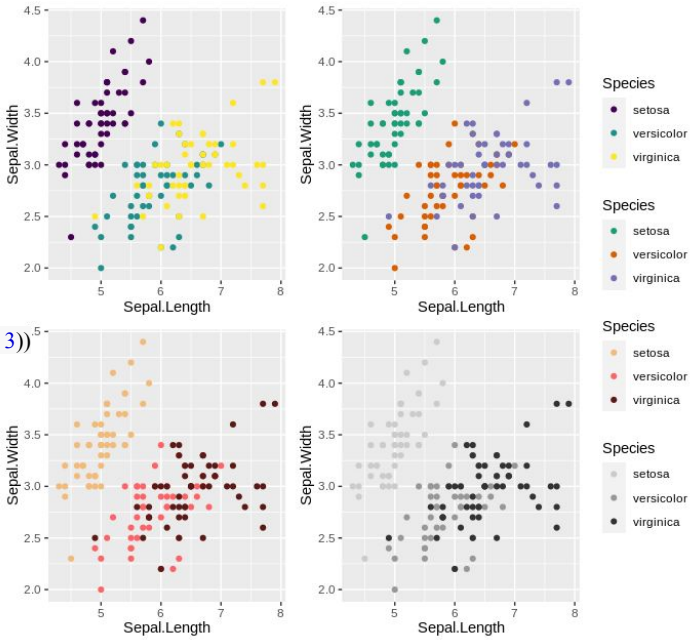
```
library(RColorBrewer) # load package
display.brewer.all() # show palettes
display.brewer.pal(n = 8, name = 'RdBu') # display specific palette
barplot(c(2,5,7), col=brewer.pal(n = 3, name = "RdBu")) # use in a plot
```



Color palettes in R

DMI

```
library(viridis)
library(wesanderson)
p <-ggplot(iris, aes(Sepal.Length, Sepal.Width))+
geom_point(aes(color = Species))
# with viridis
p1 <- p + scale_color_viridis(discrete = TRUE, option = "D")
# with rcolorbrewer
p2 <- p + scale_color_brewer(palette = "Dark2")
# with wesanderson
p3 <-p + scale_color_manual(values = wes_palette("GrandBudapest1",n = 3))
# with ggplot gray
p4 <- p + scale_color_grey(start = 0.8, end = 0.2)
library(patchwork)  # combining plots
p1 + p2 + p3 + p4 + plot_layout(ncol = 2, guides = "collect")

library(ggpubr)
ggarrange(p1, p2, p3,p4 ,
labels = c("A", "B", "C","D"),
ncol = 2, nrow = 2)
```



Color palettes in R

DMI

On your own: explore more palettes such as:

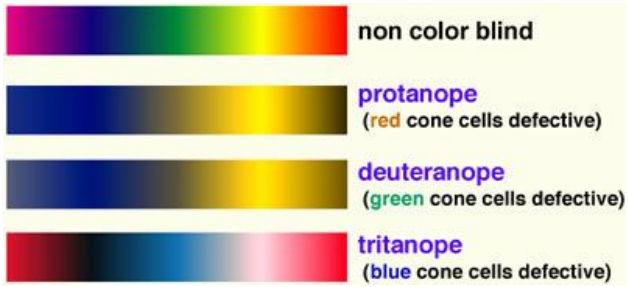
- Grey color palettes [ggplot2 package]
- Scientific journal color palettes [ggsci package]
- Wes Anderson color palettes [wesanderson package]

Common pitfalls of color use

DMI

Not designing for color-vision deficiency

How colorblind people see colors ?




```
library(RColorBrewer)
display.brewer.all(type="seq")
brewer.pal.info %>%
   head()
```

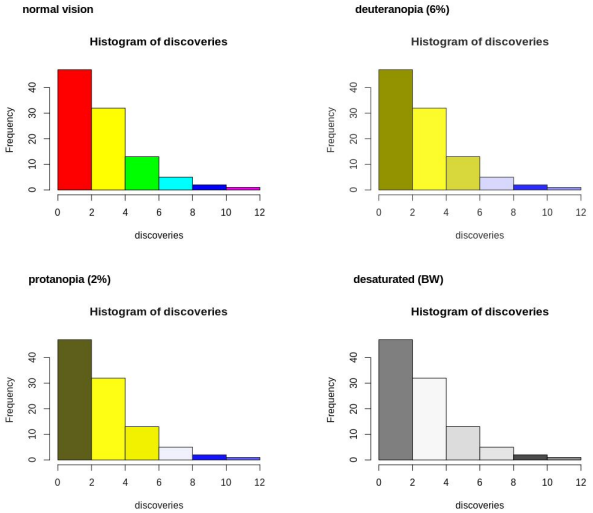
	maxcolors	category	colorblind
BrBG	11	div	TRUE
PiYG	11	div	TRUE
PRGn	11	div	TRUE
PuOr	11	div	TRUE
RdBu	11	div	TRUE
RdGy	11	div	FALSE

```
display.brewer.all(colorblindFriendly=TRUE)
```

Common pitfalls of color use

Not designing for color-vision deficiency

```
rainbowcols <- rainbow(6)
hist(discoveries,  col = rainbowcols)
library(colorBlindness)
cvdPlot(hist(discoveries,  col =
rainbowcols))
```



Differences between friendly and unfriendly graphics

DMI

Friendly	Unfriendly
words spelled out, no mysterious encodings	many abbreviations requiring the reader to decode them
words run left to right as normal	words run vertically or several directions
little messages to help explain data	graphic requires repeated reference to scattered text in some narrative at some distance from the graphic
labels are on the graphic eliminating a separate legend-- a legend pattern that follows a logical pattern	obscure codings require consulting legend repeatedly, e.g. elaborate, encoding shadings, cross hatching and color codes
graphic attracts viewer, provokes curiosity, every visual characteristic has meaning	graphic is full of chartjunk
colors are chosen so that those with color blindness can make sense of the graphic (blue is best)	design insensitive to color blindness (red and green)
type is clear using upper and lower case, simple font	Type is clotted, and in all caps, elaborated fonts

The Visual Display of Quantitative Information
E.R Tufte

The top 10 worst graphs

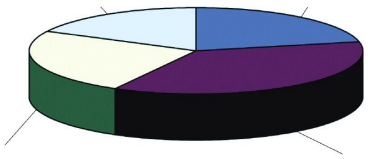
https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The top 10 worst graphs

DMI

Distribution of All TFBS Regions

What's wrong with this one?



866 Total TFBS Regions

Figure 1 Classification of TFBS Regions

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The top 10 worst graphs

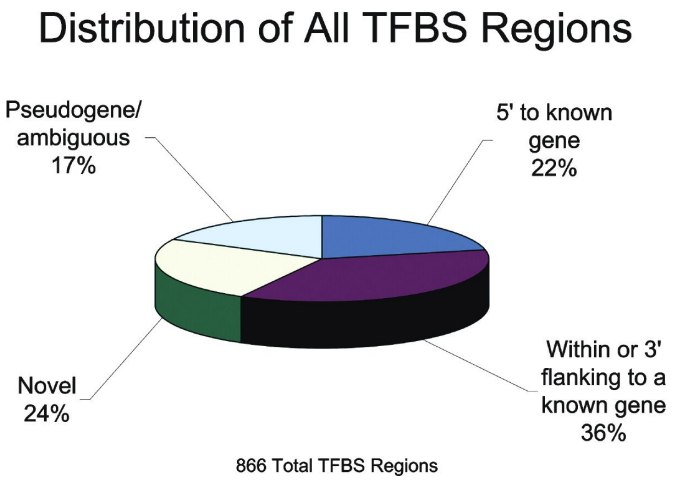


Figure 1 Classification of TFBS Regions

What's wrong with this one?

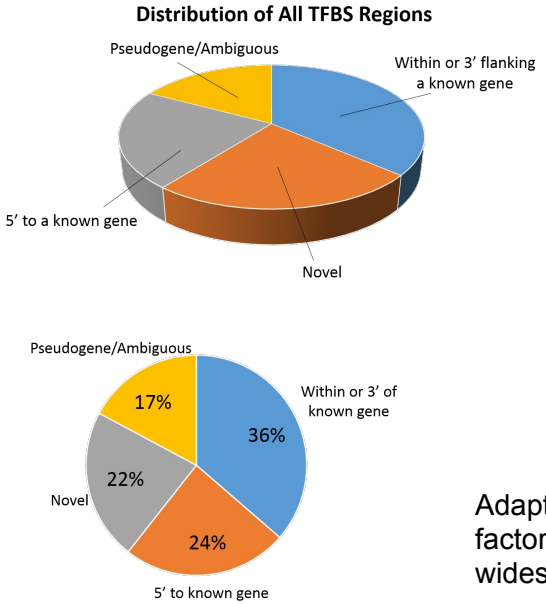
The 3D rendering of this pie chart is gratuitous. Pie charts are bad, as humans are notoriously poor at comparing areas. The color is gratuitous, too. Any graph that is meaningful only if the numbers are also cited must be viewed as a failure.

What should have been done?

The authors could have just cited the numbers. Alternatively, a bar plot (without gratuitous 3D) wouldn't be unreasonable.

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

Pie charts



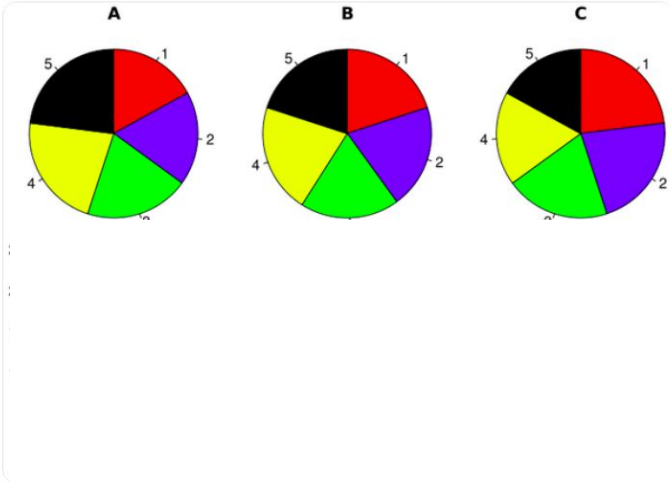
Within or 3' of known gene	36%
5' to known gene	24%
Novel	22%
Pseudogene/ambiguous	17%

Adapted from Cawley S, et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116:499-509, Figure 1

Pie charts

DMI

This is why pie charts = lie charts



What's wrong with this one?

As stated by the help file for the pie function:

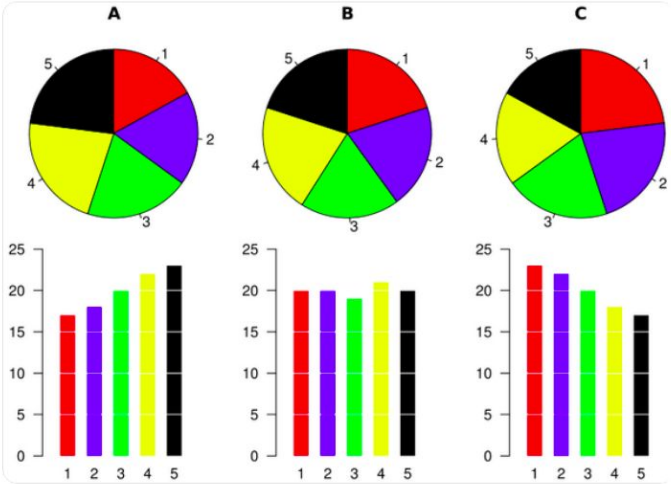
“Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.”

https://twitter.com/MonaChalabi/status/527121946073632768

Pie charts

DMI

This is why pie charts = lie charts



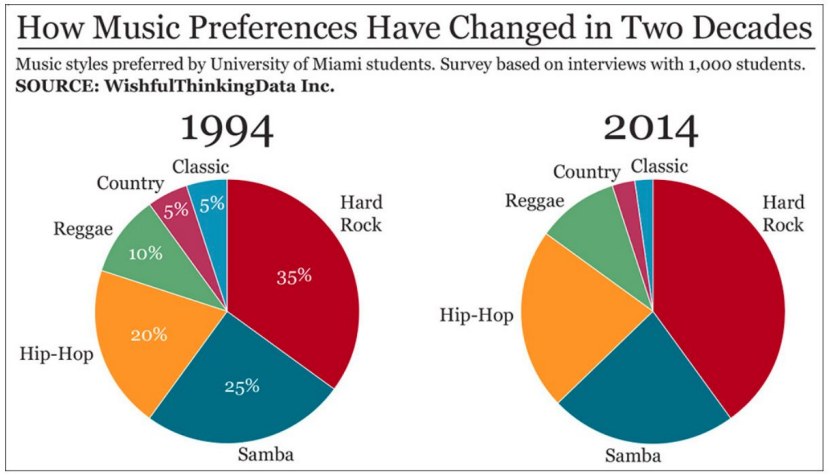
What's wrong with this one?

As stated by the help file for the pie function:

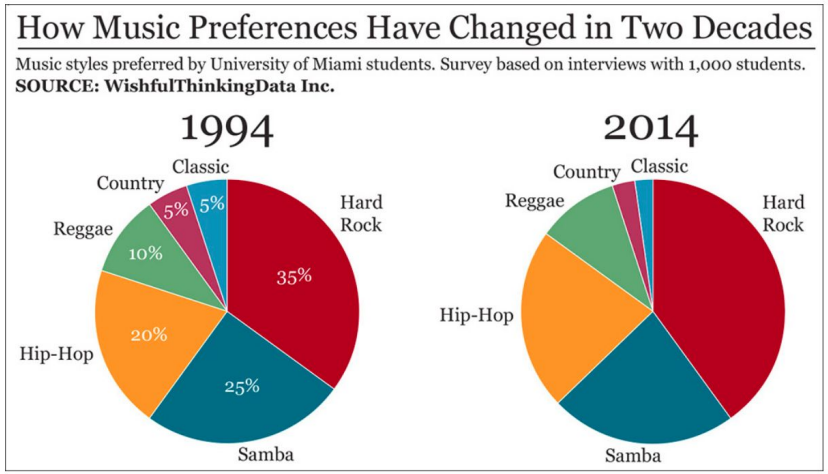
"Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data."

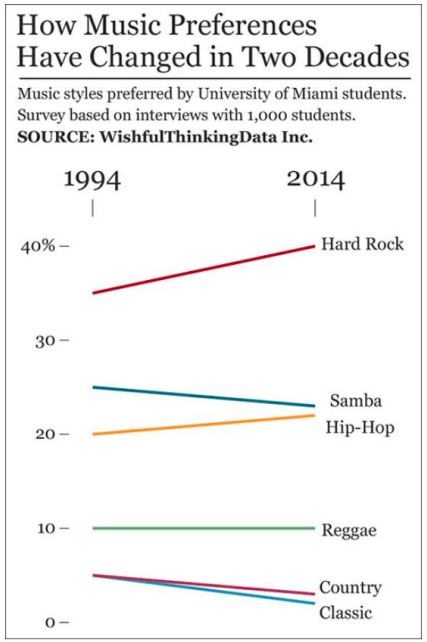
https://twitter.com/MonaChalabi/status/527121946073632768

Pie charts

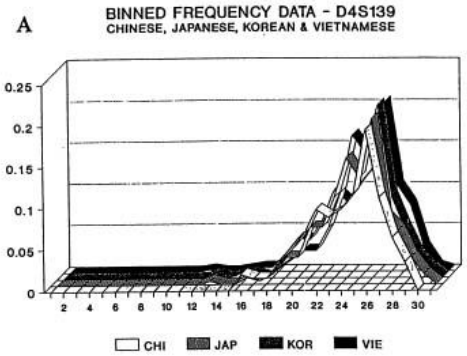

How Music Preferences Have Changed in Two Decades
Music styles preferred by University of Miami students. Survey based on interviews with 1,000 students.
SOURCE: WishfulThinkingData Inc.
1994
Classic
Country
5%
5%
Hard Rock
Reggae
10%
35%
20%
Hip-Hop
25%
Samba
2014
Country
Classic
Reggae
Hard Rock
Hip-Hop
Samba

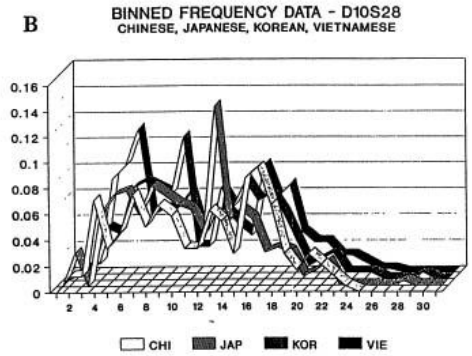
Pie charts


How Music Preferences Have Changed in Two Decades
Music styles preferred by University of Miami students. Survey based on interviews with 1,000 students.
SOURCE: WishfulThinkingData Inc.
1994
Classic
Country
5%
5%
Hard Rock
Reggae
10%
35%
Hip-Hop
20%
25%
Samba
2014
Country
Classic
Reggae
Hard Rock
Hip-Hop
Samba


How Music Preferences Have Changed in Two Decades
Music styles preferred by University of Miami students. Survey based on interviews with 1,000 students.
SOURCE: WishfulThinkingData Inc.
1994
2014
40% –
30 –
20 –
10 –
0 –
Hard Rock
Samba
Hip-Hop
Reggae
Country
Classic

The top 10 worst graphs

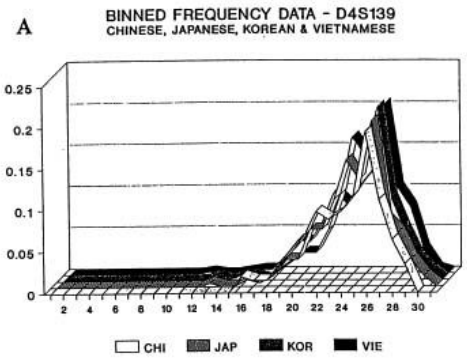




What's wrong with this one?

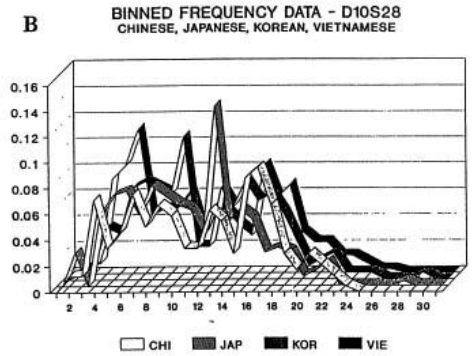
https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The top 10 worst graphs



What's wrong with this one?

Curves rendered as ribbons? The 3-dimensional rendering of the curves is entirely gratuitous.



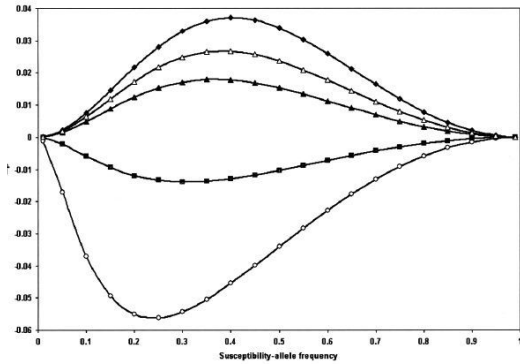
What should have been done?

It's difficult to display multiple curves simultaneously and ensure that the individual curves may be seen. Colors would be nice, but if color is not allowed, four different line types (solid, dashed, dotted, dash-dotted) might work.

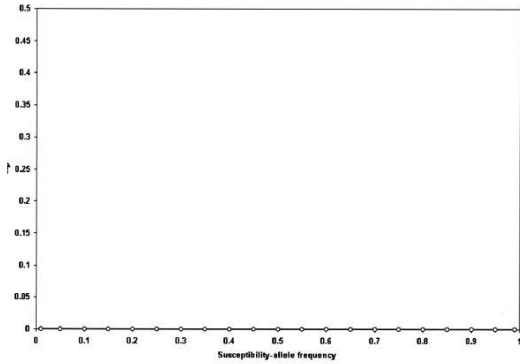
https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The top 10 worst graphs

C



D

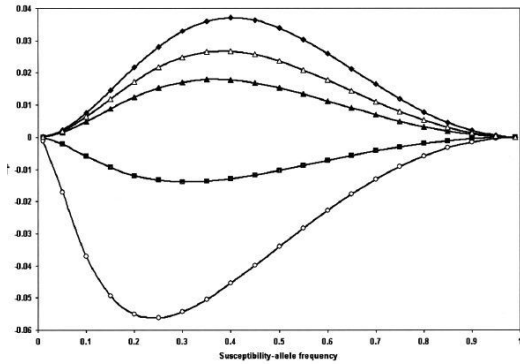


What's wrong with this one?

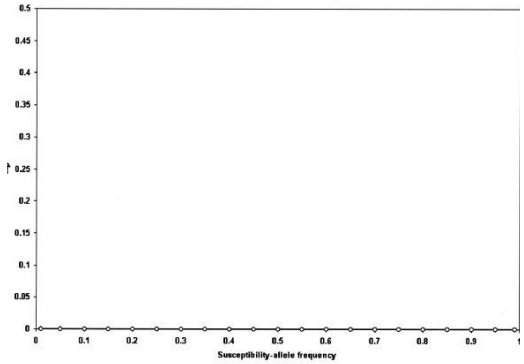
https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The top 10 worst graphs

C



D



What's wrong with this one?

This figure spans two pages. Panel D is most interesting; it takes a while to identify that there is any information there at all.

What should have been done?

Figure 1D could have been discarded completely, and the whole figure shouldn't take up more than half a page.

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The top 10 worst graphs

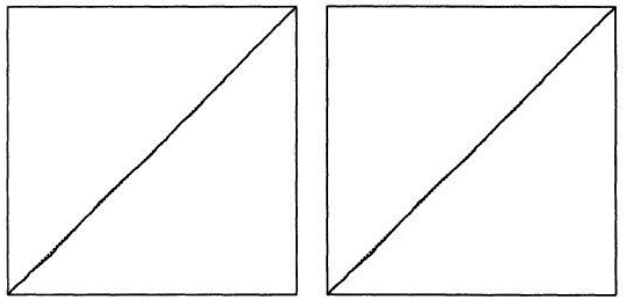


Figure 1. SRQ Plots of T_i/T_n (Vertical Axes) Against i/n (Horizontal Axes) for the Gibbs Sampler (a) and an Alternating Gibbs/Independence Sampler (b) for the Pump Failure Data Based on Runs of Length 5,000. Lines through the origin with unit slope are shown dashed; axis ranges are from 0 to 1 for all axes.

What's wrong with this one?

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The top 10 worst graphs

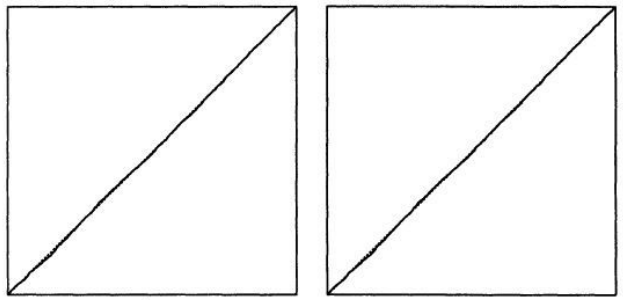


Figure 1. SRQ Plots of T_i/T_n (Vertical Axes) Against i/n (Horizontal Axes) for the Gibbs Sampler (a) and an Alternating Gibbs/Independence Sampler (b) for the Pump Failure Data Based on Runs of Length 5,000. Lines through the origin with unit slope are shown dashed; axis ranges are from 0 to 1 for all axes.

What's wrong with this one?

it contains almost no information. This graphic is totally uninformative.

What should have been done?

Plot percent differences, or just say the results are indistinguishable.

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The top 10 worst graphs

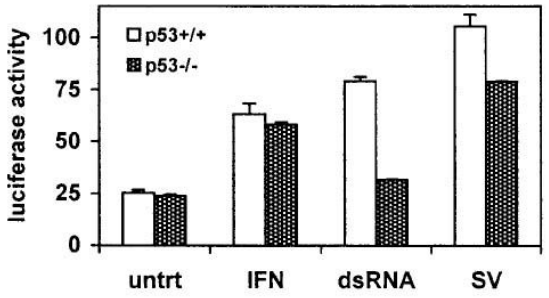


FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNA, and virus. Cells (6×10^5 HCT 116) were seeded in 32-mm plates and allowed to attach overnight. Cells were transfected with 500 ng of pGL3/ISG15-Luc, 50 ng of pRL null (Promega), and 450 ng of pcDNA3 for carrier DNA by using Lipofectamine Plus (Life Technologies) following the manufacturer's instructions. Twenty-four hours posttransfection, the medium was aspirated and replaced with medium containing either 1,000 U of IFN-α/ml, 50 μg of dsRNA/ml, or Sendai virus (multiplicity of infection, 10). Cells were incubated for 12 h and then lysed, and luciferase assays were performed. Luciferase activity was assessed on 20 μl of each lysate as directed by the supplier (Dual Luciferase Kit, Promega) using a TD 20/20 luminometer (Turner Designs). Luciferase activity is presented as the ratio of firefly activity to renilla activity to control for differences in transfection efficiency. Each data point is the mean of triplicate samples ± the standard error; the data presented are representative of four independent experiments.

What's wrong with this one?

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The top 10 worst graphs

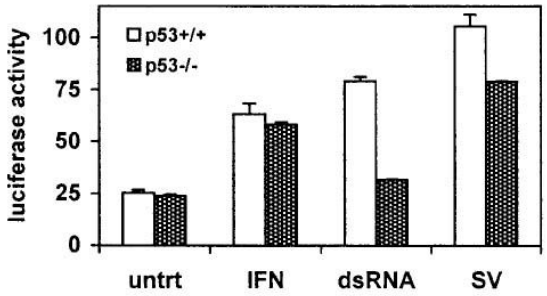


FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNA, and virus. Cells (6×10^5 HCT 116) were seeded in 32-mm plates and allowed to attach overnight. Cells were transfected with 500 ng of pGL3/ISG15-Luc, 50 ng of pRL null (Promega), and 450 ng of pcDNA3 for carrier DNA by using Lipofectamine Plus (Life Technologies) following the manufacturer's instructions. Twenty-four hours posttransfection, the medium was aspirated and replaced with medium containing either 1,000 U of IFN-α/ml, 50 μg of dsRNA/ml, or Sendai virus (multiplicity of infection, 10). Cells were incubated for 12 h and then lysed, and luciferase assays were performed. Luciferase activity was assessed on 20 μl of each lysate as directed by the supplier (Dual Luciferase Kit, Promega) using a TD 20/20 luminometer (Turner Designs). Luciferase activity is presented as the ratio of firefly activity to renilla activity to control for differences in transfection efficiency. Each data point is the mean of triplicate samples ± the standard error; the data presented are representative of four independent experiments.

What's wrong with this one?

The bars and little antennae represent just three data points each.

What should have been done?

With just three data points in each group, why not just show the data as dots? You could also include line segments at the averages and even confidence intervals...all this in the same amount of space and with less ink.

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The top 10 worst graphs

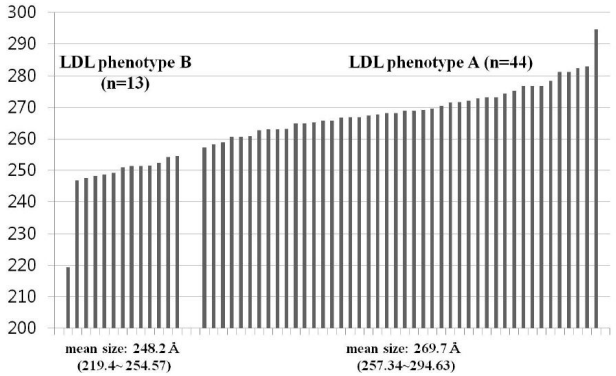


Fig. 1. Distribution of low-density lipoprotein (LDL) particle size in all stuc subjects (LDL phenotypes A and B). *LDL phenotype A group* (mean size: 269 Å, n = 44), subjects with buoyant-mode profiles [peak LDL particle diameter ≥ 2(Å] including intermediate LDL subclass pattern [256 Å ≤ peak LDL particle diamet ≤ 263 Å]; *LDL phenotype B group* (mean size: 248.2 Å, n = 13), subjects wi dense-mode profiles [peak LDL particle diameter ≤ 255 Å]

What's wrong with this one?

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The top 10 worst graphs

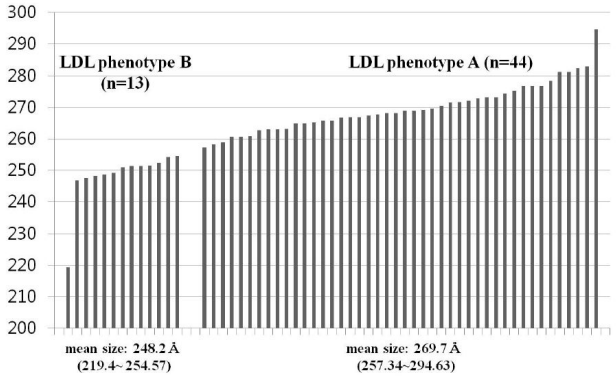


Fig. 1. Distribution of low-density lipoprotein (LDL) particle size in all study subjects (LDL phenotypes A and B). *LDL phenotype A group* (mean size: 269.7 Å, n = 44), subjects with buoyant-mode profiles [peak LDL particle diameter ≥ 264 Å] including intermediate LDL subclass pattern [256 Å ≤ peak LDL particle diameter ≤ 263 Å]; *LDL phenotype B group* (mean size: 248.2 Å, n = 13), subjects with dense-mode profiles [peak LDL particle diameter ≤ 255 Å]

What's wrong with this one?

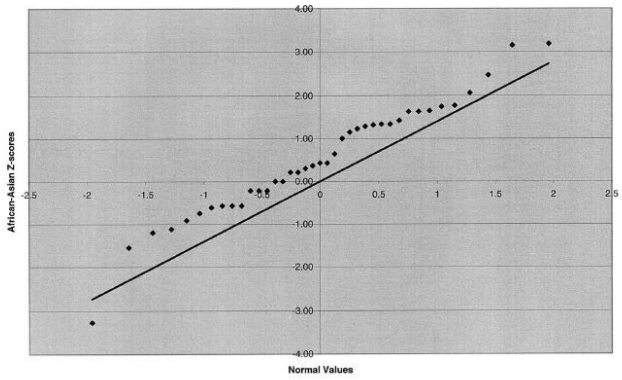
While this is better than having an antenna chart for each group, using a separate bar for each individual is unnecessary and hard to read.

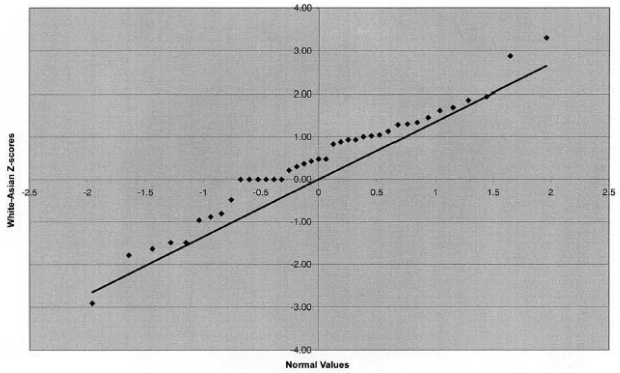
What should have been done?

It would be much better to use dots rather than bars for the individual values.

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

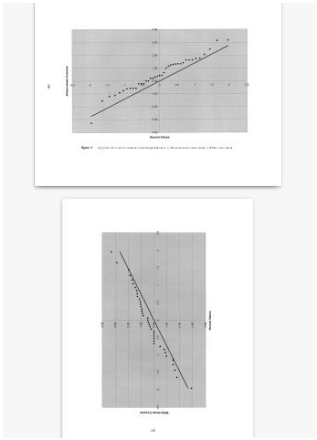
The top 10 worst graphs





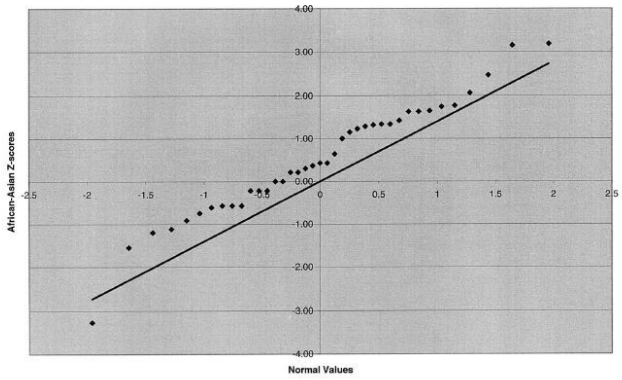
What's wrong with this one?

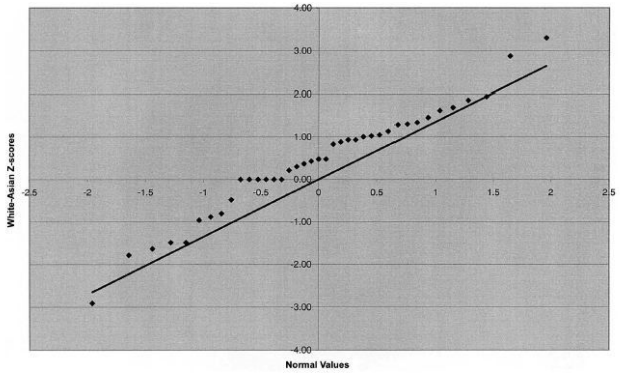
se



https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The top 10 worst graphs





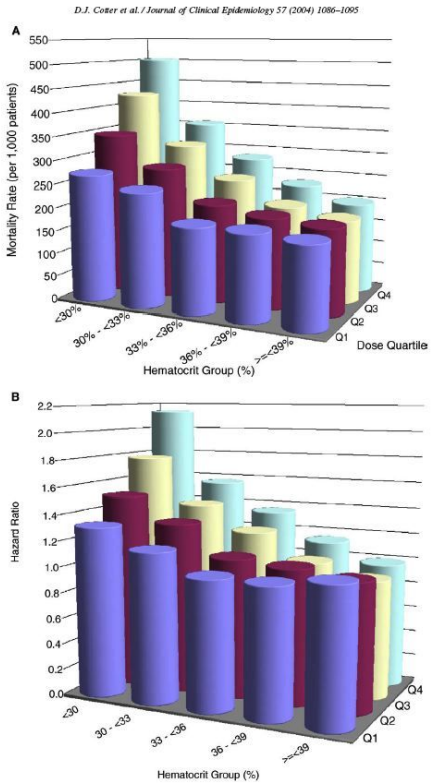
What's wrong with this one?

There's always a lot of wasted space in QQ plots, but these are particularly bad: the figure takes two full pages, and that gray background! Wasted spaced and ink.

What should have been done?

Better dot plots showing the values in all groups, side-by-side, with a white background, using one-quarter of the page.

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

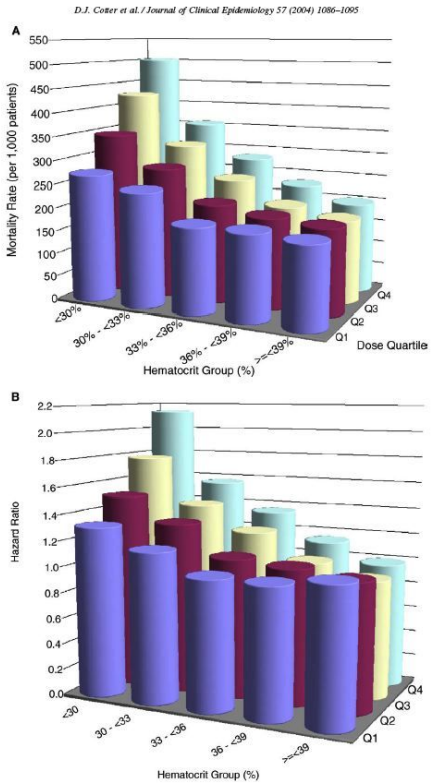


The top 10 worst graphs

DMI

What's wrong with this one?

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/



The top 10 worst graphs

What's wrong with this one?

The perspective makes it difficult to compare the heights of the cylinders, as the vertical scale changes from front to back. Also, this is a lot of space (and color ink) to convey very little information.

What should have been done?

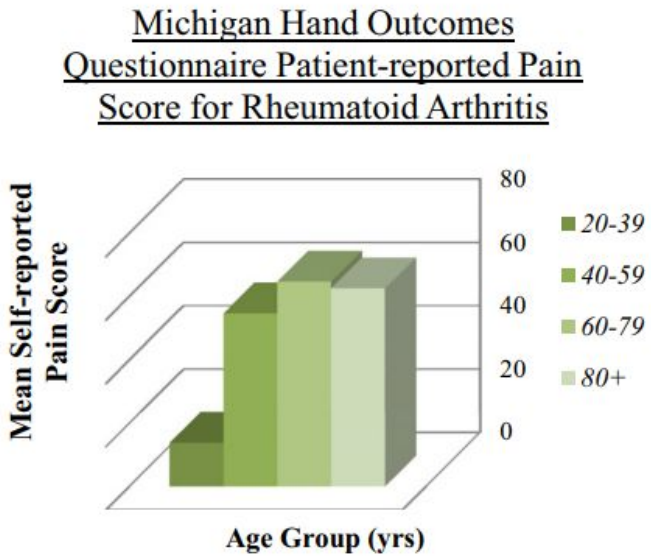
it's tricky to fix this figure; one might try four superposed lines, or maybe facets?

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

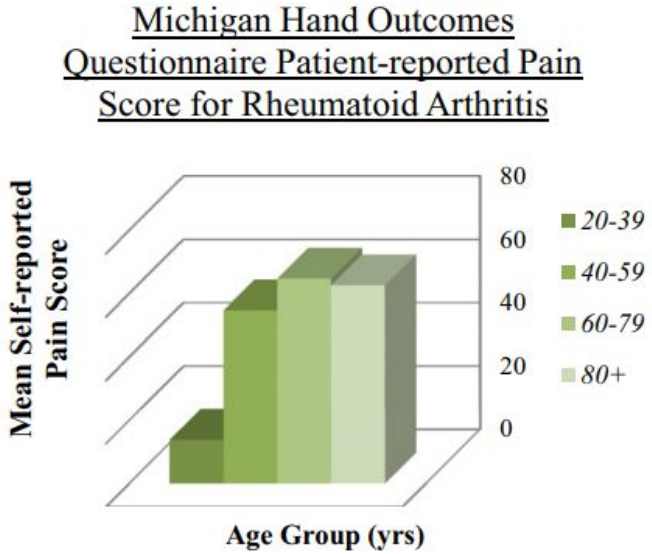
DMI

What's wrong with this one?



https://linkinghub.elsevier.com/retrieve/pii/S0363-5023(11)01653-4

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter



What's wrong with this one?

The perspective makes it difficult to compare the heights

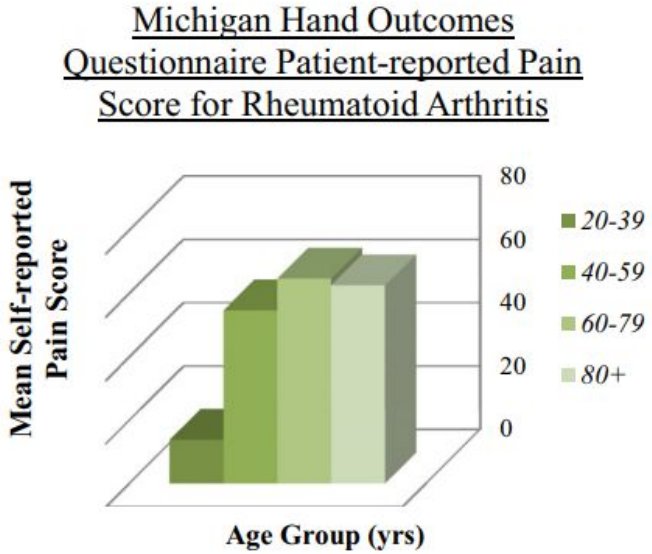
The color scheme of this graph, chosen by the computer, is monochromatic, which some readers would find challenging to distinguish

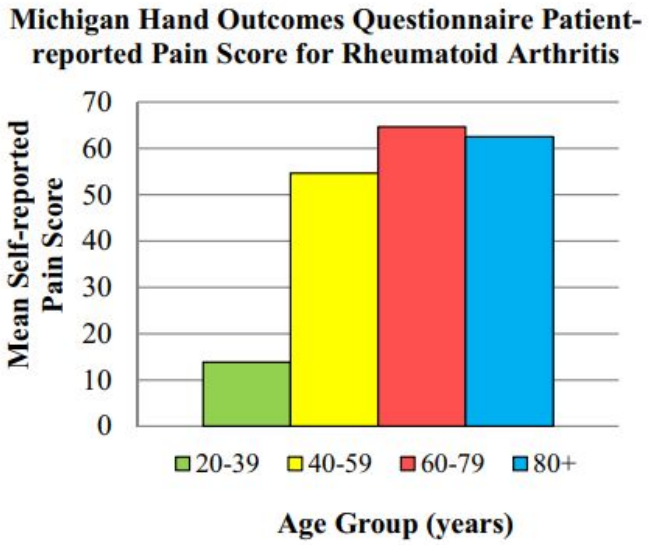
The use of underlining, italicizing, and boldface type is overwhelming

What should have been done?

https://linkinghub.elsevier.com/retrieve/pii/S0363-5023(11)01653-4

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

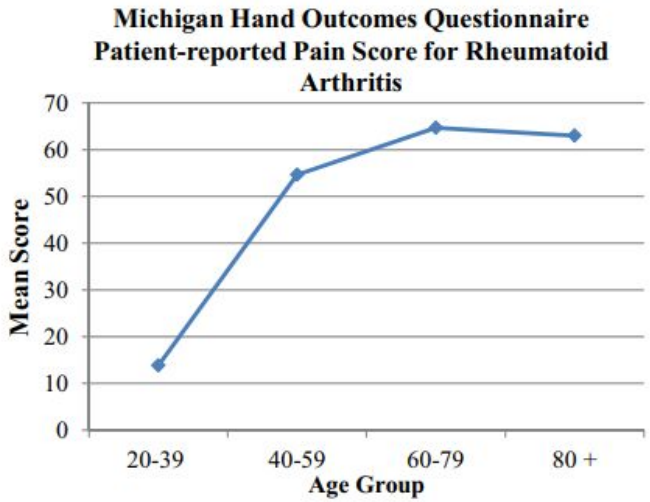




https://doi.org/10.1016/j.jhsa.2011.12.041

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

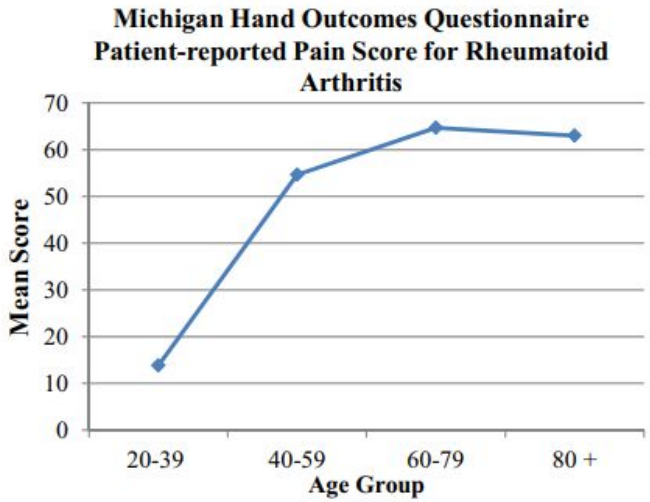
What's wrong with this one?



https://linkinghub.elsevier.com/retrieve/pii/S0363-5023(11)01653-4

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

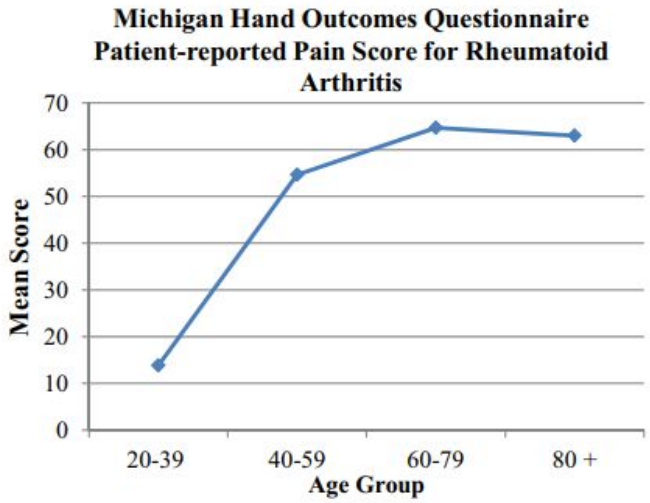
What's wrong with this one?

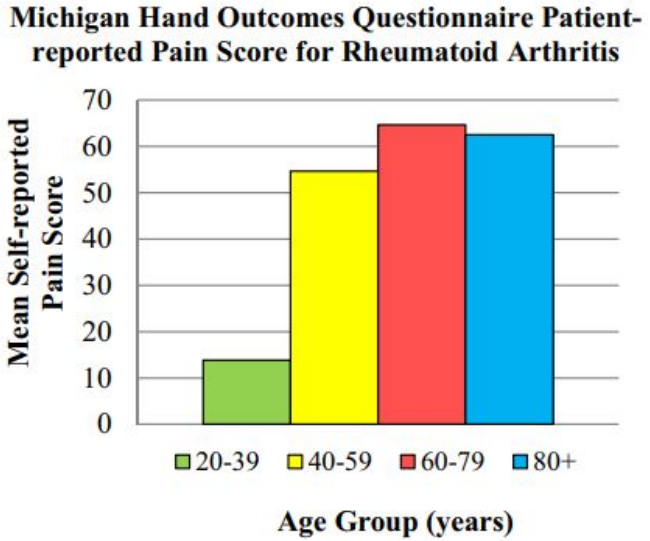


Computer-made graphs can also contain errors that mislead readers: for instance, connecting discrete data points (eg, a series of average measurements taken from a group of patients) with a continuous line. The connecting segments suggest that there are values between age groups that fall on the lines, when in fact the author cannot know this.

https://linkinghub.elsevier.com/retrieve/pii/S0363-5023(11)01653-4

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

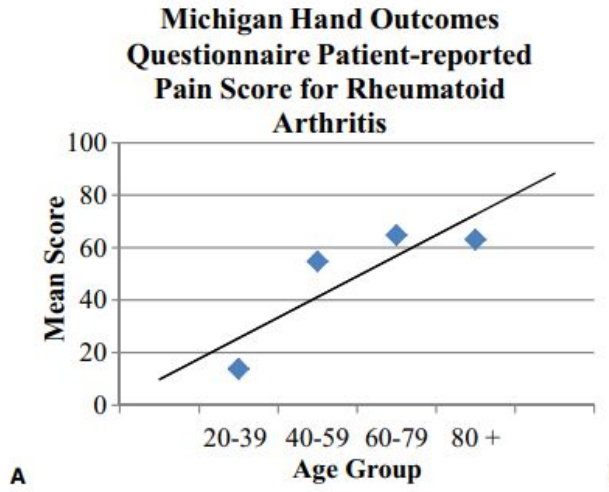




https://linkinghub.elsevier.com/retrieve/pii/S0363-5023(11)01653-4

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

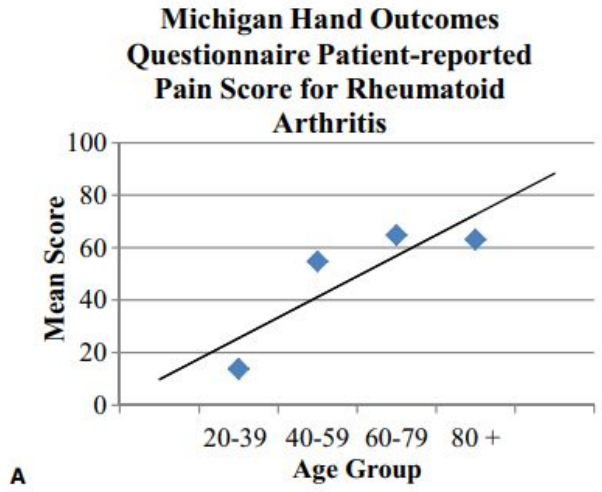
What's wrong with this one?



https://linkinghub.elsevier.com/retrieve/pii/S0363-5023(11)01653-4

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

What's wrong with this one?

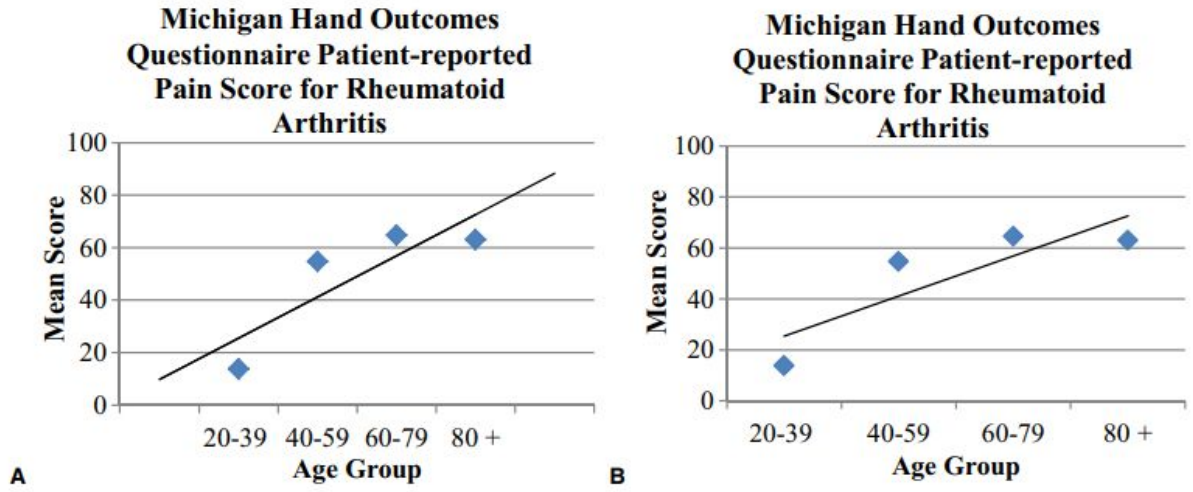


Another graphical mistake is extending a regression line beyond the domain of the data.

Regression lines should not be extrapolated beyond the set of measured values

https://linkinghub.elsevier.com/retrieve/pii/S0363-5023(11)01653-4

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter



https://linkinghub.elsevier.com/retrieve/pii/S0363-5023(11)01653-4

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

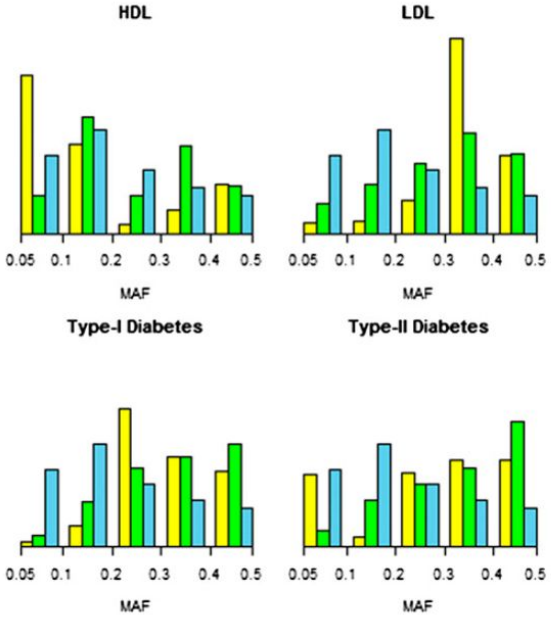
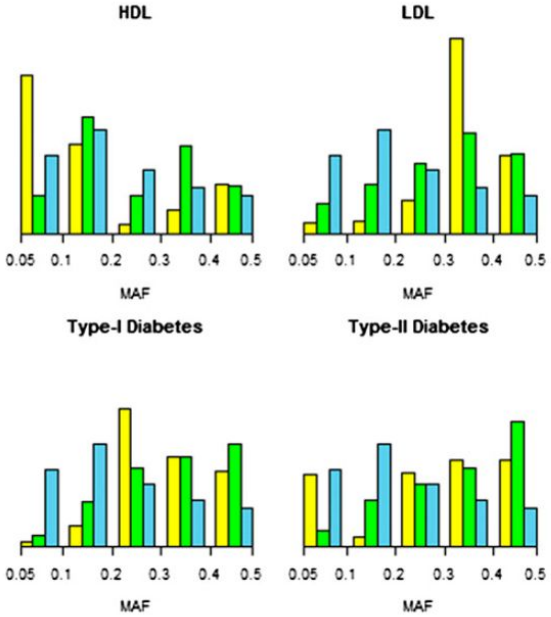


Fig. 1. Distribution of frequencies for minor alleles across an estimated number of susceptibility SNPs (yellow), observed susceptibility SNPs (green), and independent representative SNPs in the HapMap project (blue).

What is wrong with this graph?

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

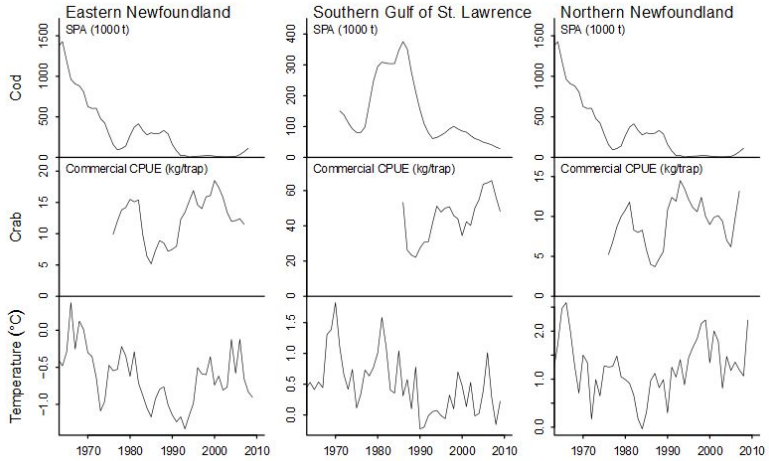


It is also important to avoid distorting the X-axis. Note in the example below that the space between 0.05 to 0.1 is the same as space between 0.1 and 0.2.

Fig. 1. Distribution of frequencies for minor alleles across an estimated number of susceptibility SNPs (yellow), observed susceptibility SNPs (green), and independent representative SNPs in the HapMap project (blue).

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

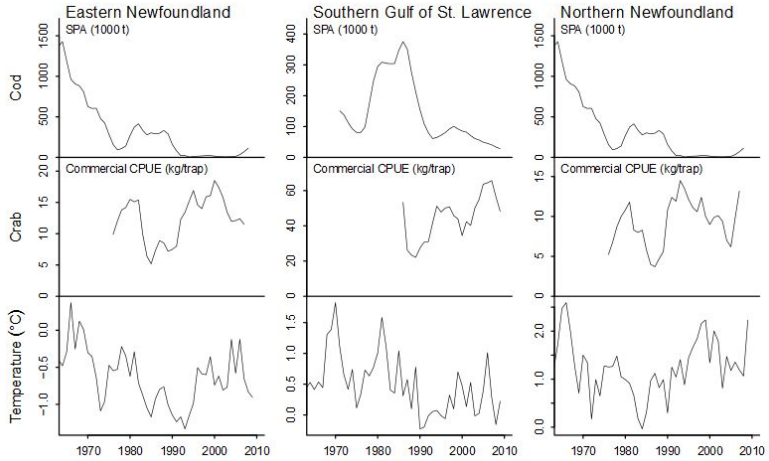


Comparison of catches of cod fish and crab across regions in relation to the variation to changes in water temperature.

What is wrong with this graph?

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

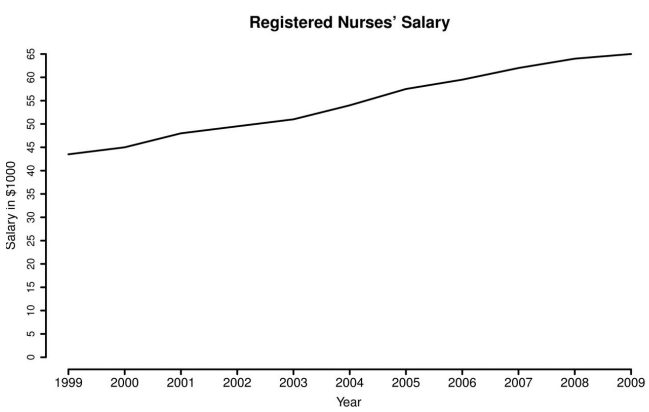


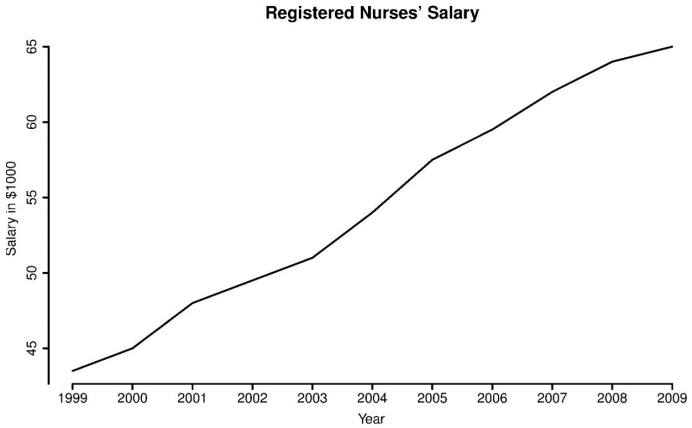
Comparison of catches of cod fish and crab across regions in relation to the variation to changes in water temperature.

The problem here is that the Y-axes are vastly different, making it hard to sort out what's really going on. Even the Y-axes for temperature are vastly different.

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

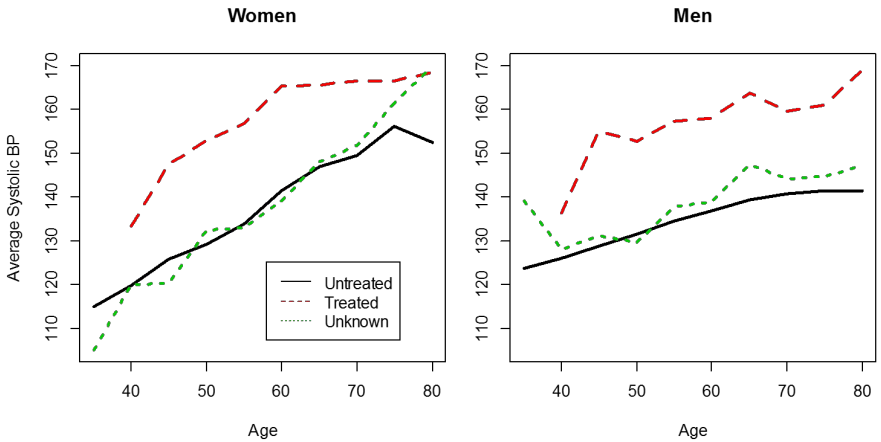




Consider the range of the Y-axis. In the examples below there is no relevant information below $40,000, so it is not necessary to begin the Y-axis at 0. The graph on the right makes more sense.

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

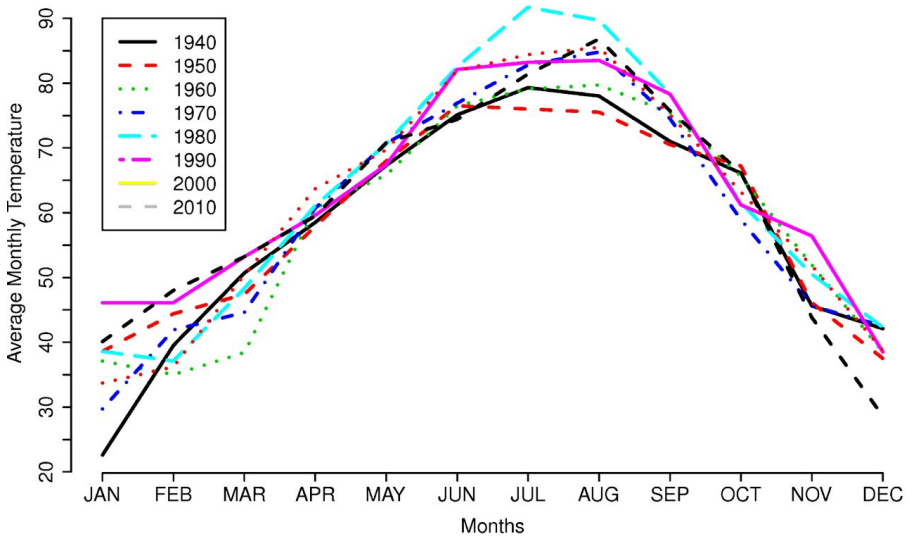
Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter



The ability to make comparisons is greatly facilitated by using the same scales for axes

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter



Avoid putting too many lines on the same chart. In the example below, the only thing that is readily apparent is that 1980 was a very hot summer

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

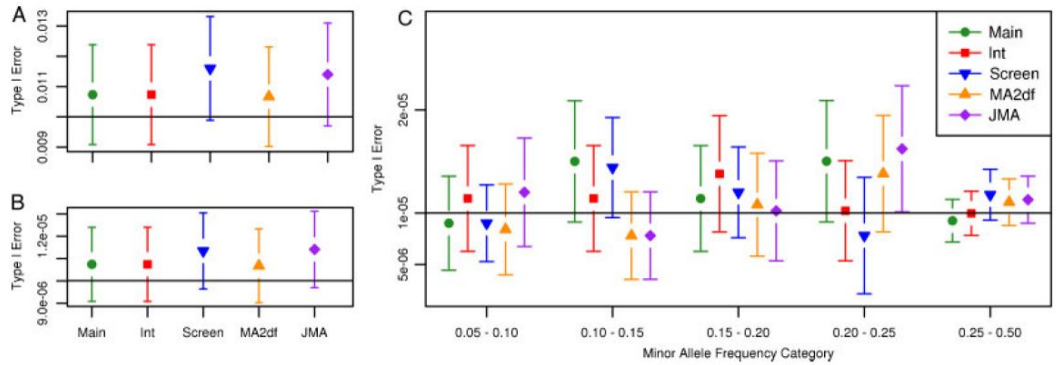
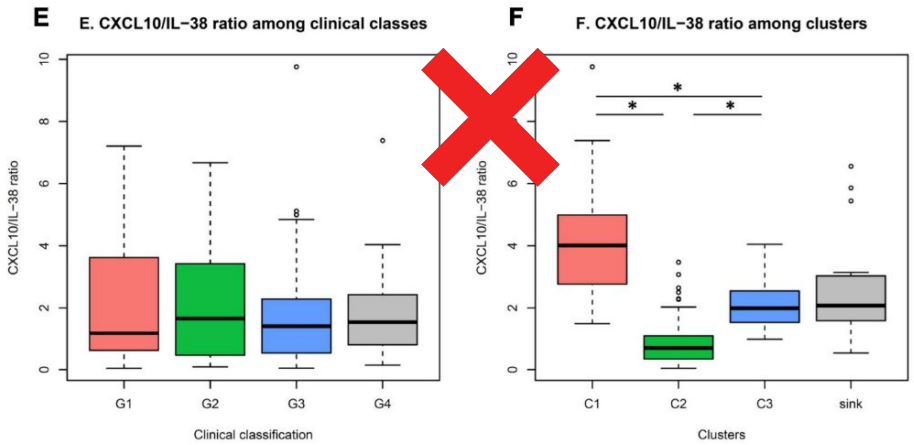


Fig. 1. (A) Empirical type I error rate in 15,000 simulations of 999 null SNPs, empirical type I error rates for (B) 14,985,000 null SNPs and (C) 14,985,000 null SNPs broken into minor allele frequency categories. A horizontal line is drawn at the theoretical type I error rate in each plot.

When multiple comparisons are being made, it is essential to use colors and symbols in a consistent way

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

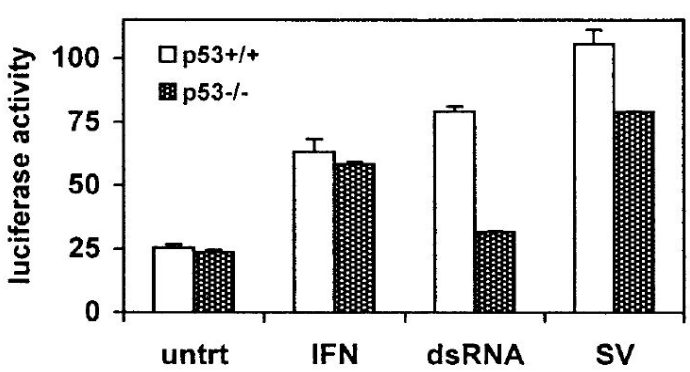
Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

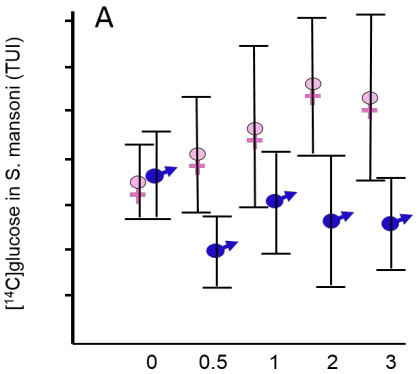


When multiple comparisons are being made, it is essential to use colors and symbols in a consistent way

https://doi.org/10.3389/fmed.2022.987182

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

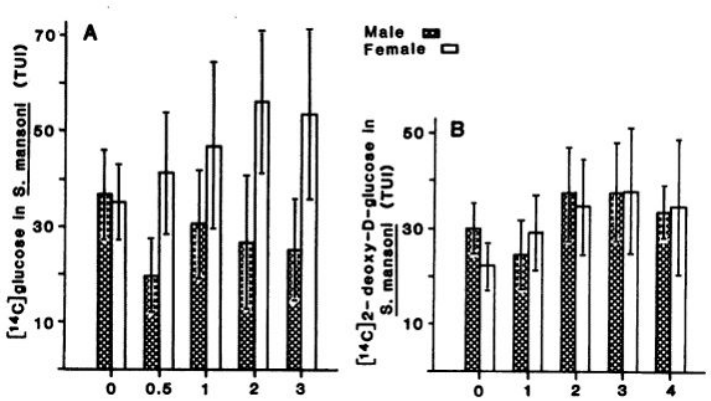


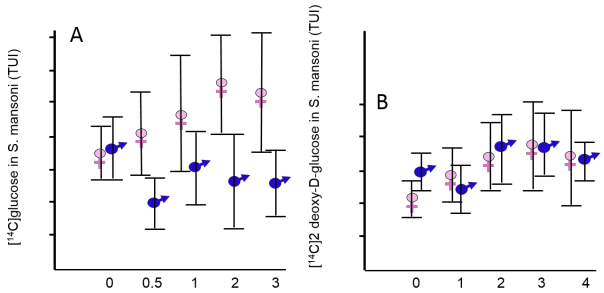


Bar charts can be problematic. The figure on the left presents means and error bars, but the error bars are misleading because they only extend in one direction. A better alternative would have been to to use full error bars with a scatter plot, as illustrated in the right.

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

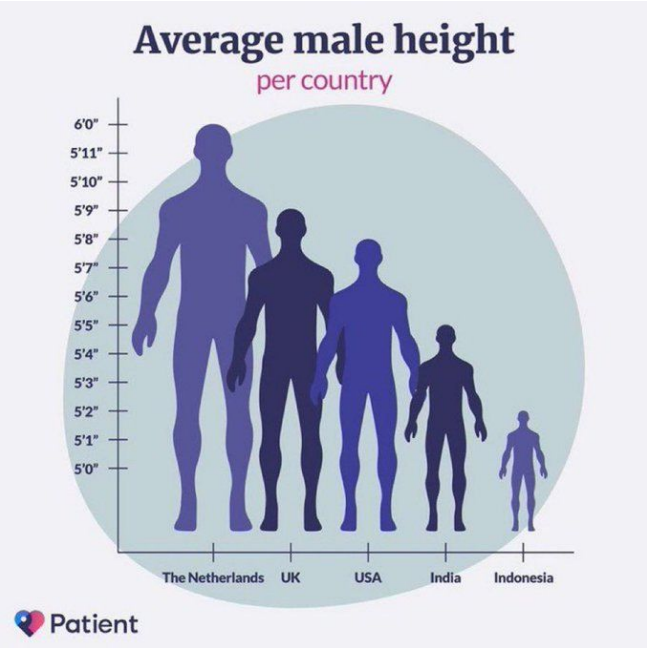




Bar graphs add ink without conveying any additional information, and they are distracting. The graph below on the left inappropriately uses bars which clutter the graph without adding anything. The graph on the right displays the same data, by does so more clearly and with less clutter.

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

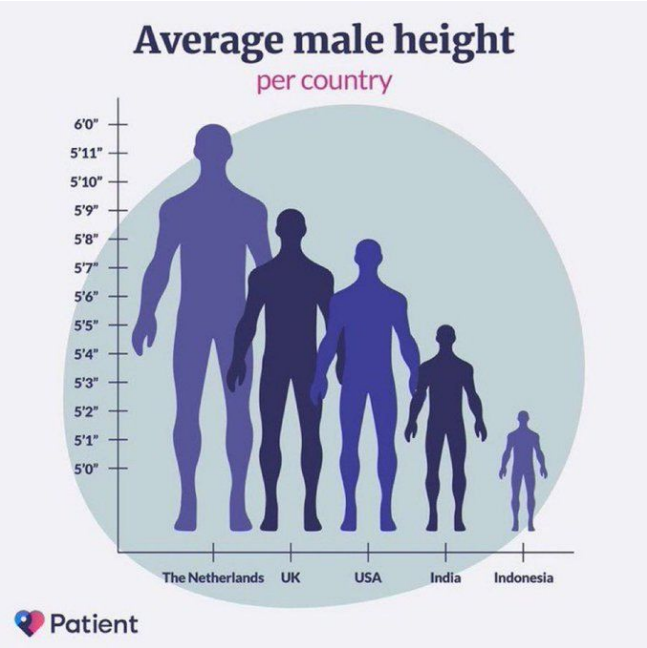
Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter



What's wrong with this one?

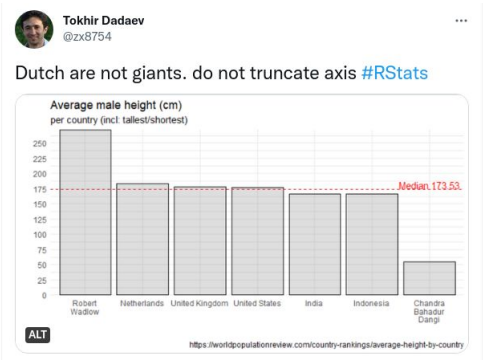
https://twitter.com/iamvladyashin/status/1602225806712995841

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter



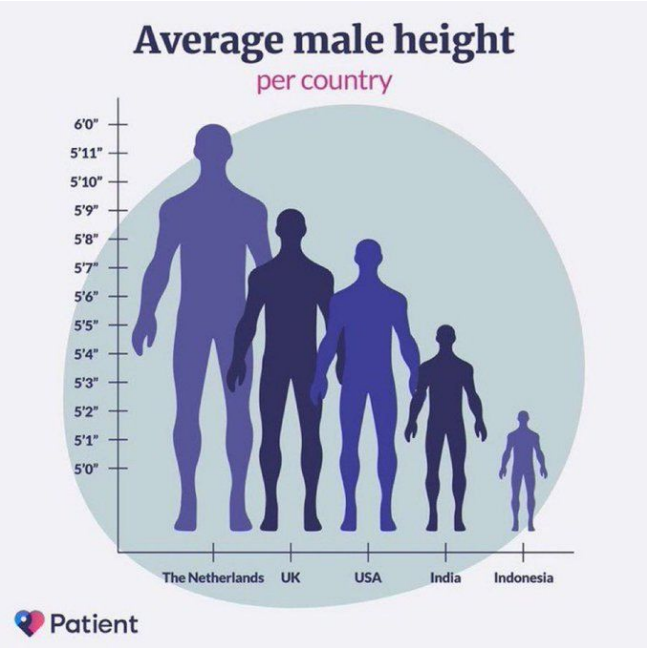
What's wrong with this one?

The truncated Y axis exaggerates height differences



https://twitter.com/iamvladyashin/status/1602225806712995841

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter



What's wrong with this one?

The truncated Y axis exaggerates height differences

To make a graph extra misleading, show data using shapes whose area is the square of a threshold number: using two-dimensional figures, scaled proportionally, turns linear differences into quadratic ones.

Doesn't use International System of Units

Doesn't indicate error estimates

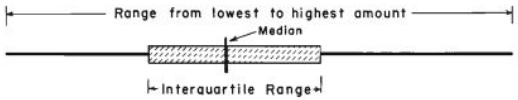
The use of color??

https://twitter.com/kareem_carr/status/1612502338547159064

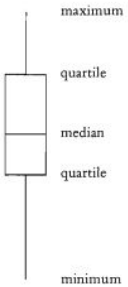
Box Plots

DMI

Mary Eleanor Spear's "range bar"



and John Tukey's "box plot"



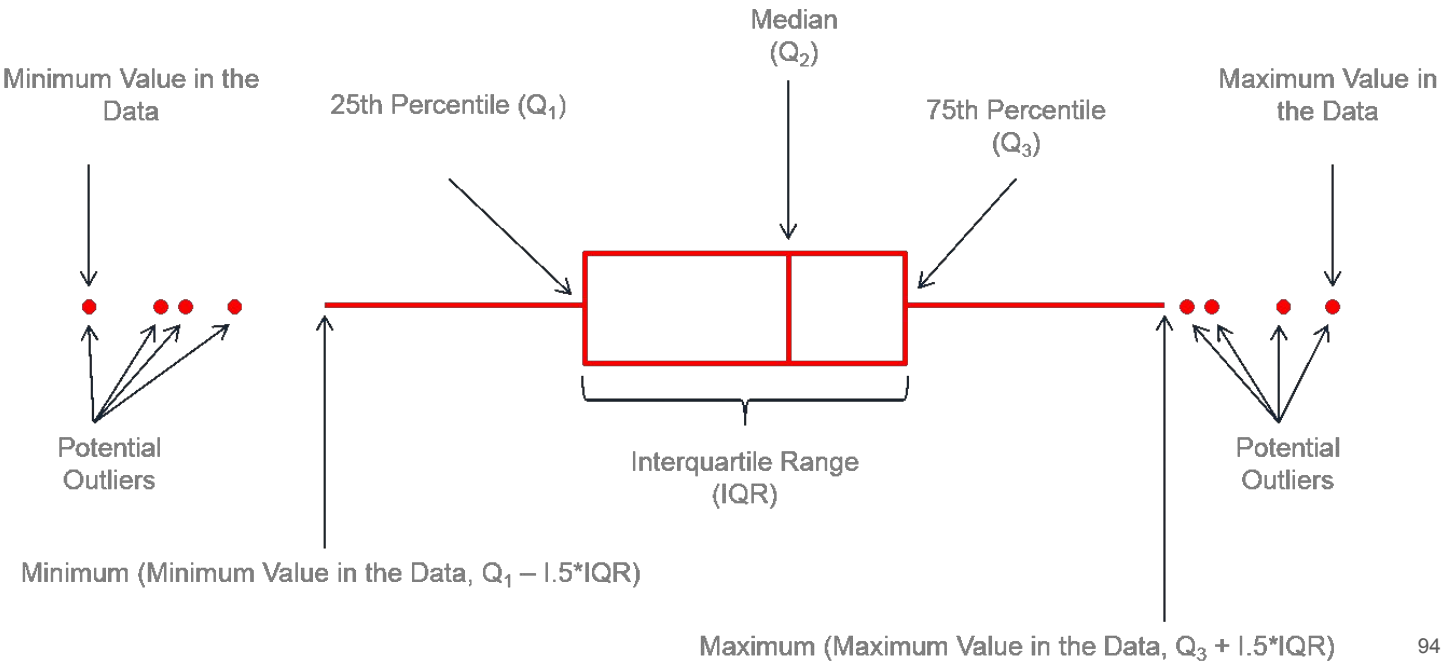
Mary Eleanor Spear, *Charting Statistics* (New York, 1952), p. 166; and John W. Tukey, *Exploratory Data Analysis* (Reading, Massachusetts, 1977).

Best for:

- **Numerical (continuous) data**
- **Distribution comparisons** across groups

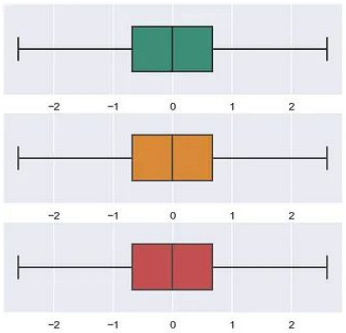
Box Plots

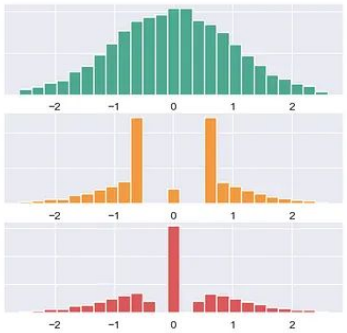
DMI

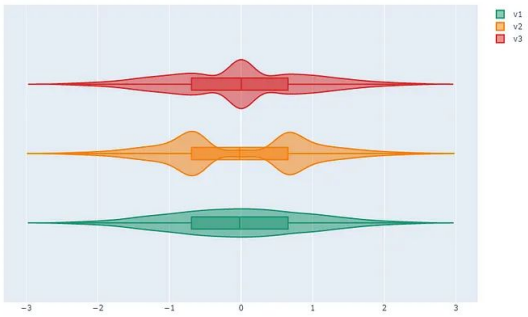


Box Plots

DMI





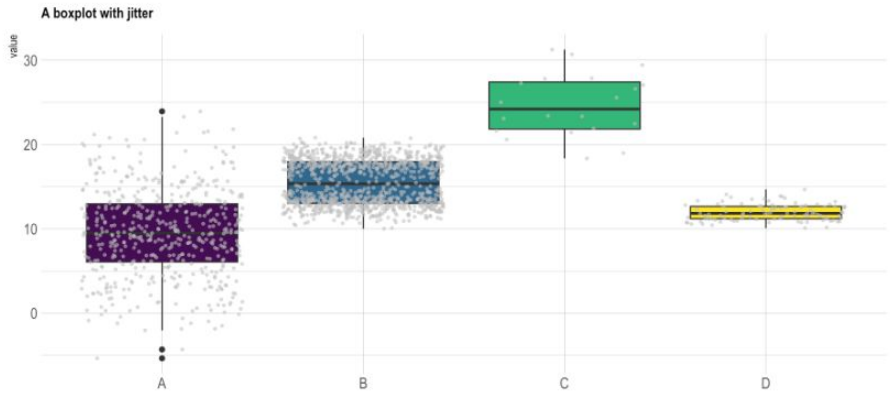


Box Plots

DMI

```
head(data)
name      value
   A  6.7307422
   A 15.3695408
   A -0.1719162
   A  2.6875983
   A 11.9881003
   A 15.8797916
```

```
# 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("")
```



Boxplot is probably the most commonly used chart type to compare distribution of several groups.

https://www.data-to-viz.com/caveat/boxplot.html

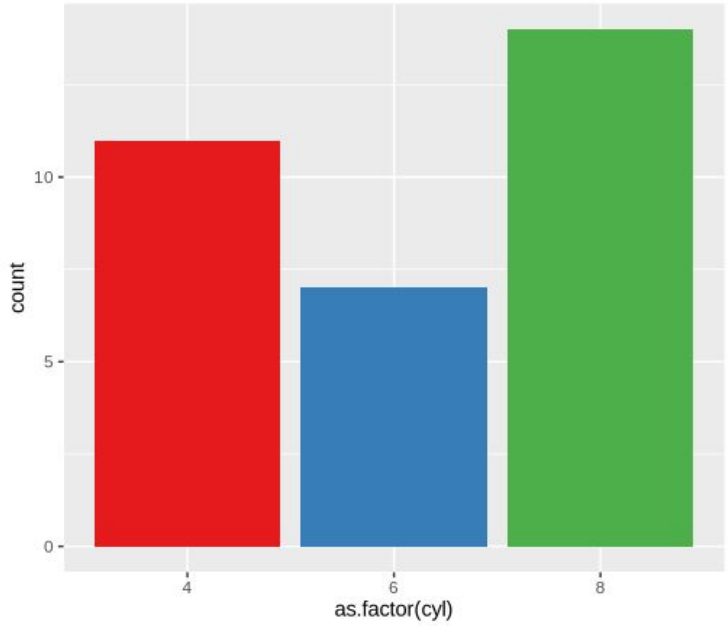
BUT: Don't use boxplots for small numbers of observations, just plot the data!

Bar Plots

DMI

```
> head(mtcars)
                   mpg cyl disp
Mazda RX4         21.0   6  160 1
Mazda RX4 Wag     21.0   6  160 1
Datsun 710        22.8   4  108
Hornet 4 Drive    21.4   6  258 1
Hornet Sportabout 18.7   8  360 1
Valiant           18.1   6  225 1
>
```

```
ggplot(mtcars, aes(x=as.factor(cyl),
fill=as.factor(cyl) )) +
  geom_bar( ) +
  scale_fill_brewer(palette = "Set1") +
    theme(legend.position="none")
```



Bar plots

Bar plots have many problems:

- High ink to information ratio
- Error bars cause perception errors
- Can only show one-sided confidence intervals well
- Thick bars reduce the number of categories that can be shown
- Labels on vertical bar charts may be difficult to read

Dots plots are almost always better

Consider multi-panel side-by-side display for comparing several contrasting or similar cases

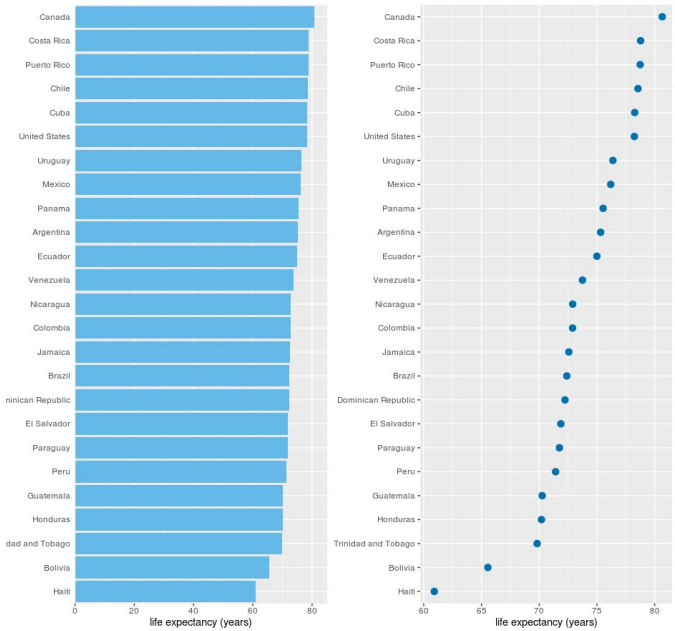
Use same scales for both x- and y-axes across different panels

Consider ordering categories by values represented, for more accurate perception

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

Tips

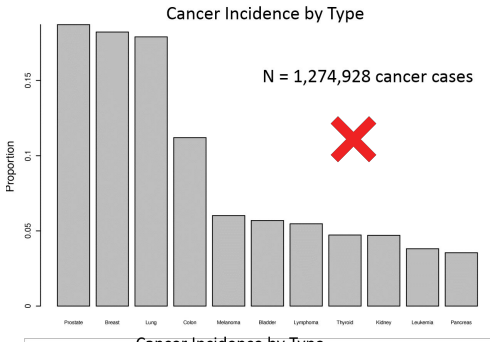
DMI

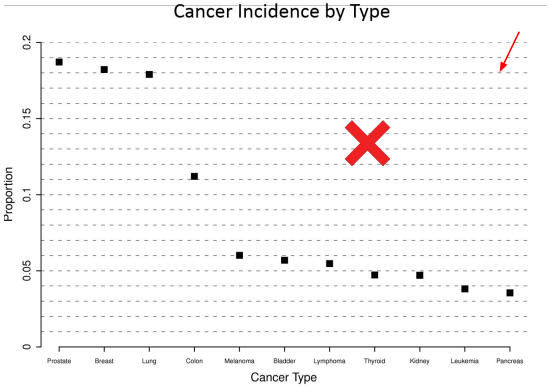


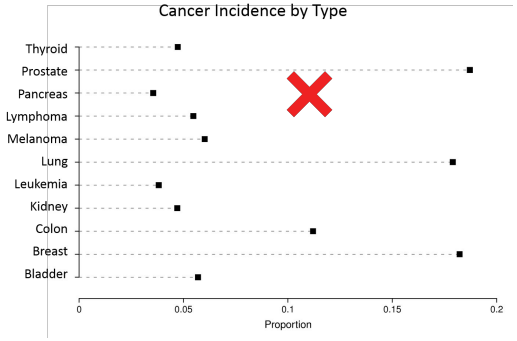
Visualizing amounts: dot plots

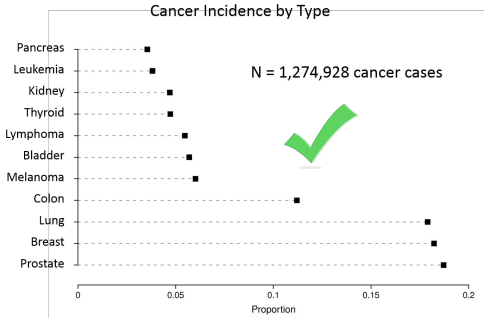
```
library(gapminder)

df_Americas <- gapminder %>% filter(year == 2007,
continent == "Americas")

ggplot(df_Americas, aes(x = lifeExp,
          y =  reorder(country, lifeExp))) +
    geom_point(color = "#0072B2", size = 3) +
    scale_x_continuous(
          name = "life expectancy (years)",
        limits = c(59.7, 81.5)    ) +
    scale_y_discrete(name = NULL, expand = c(0,
 0.5))  + theme_minimal()
```

Tips









http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

Tips

DMI

Reduce Ink to information ratio

Bar charts not appropriate to display means (high ink to information ratio)

Adjust x- and y-axes if needed

Consider using a table instead of a plot

Show multiple types of information in same figure

Use different size dots to represent sample size

Use 'heat map' or 'hues' to represent different levels of a variable

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

Tips

DMI

- Consistent use of x and y-axes across multiple panels
- Carefully consider the inclusion of "0" in your axis: sometimes, it is essential to include 0, often, inclusion of 0 is not necessary
- Consider using a log scale when it is important to understand percent change of multiple factors
- Consistent use of colors for different categories
- Consistent use of fonts, line widths, box sizes, etc., to avoid distortion
- With few categories, a single figure may facilitate comparisons; with many categories, consider multiple panels

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

Tips

DMI

- In general, avoid bar plots
- Avoid chart junk and the use of too much ink relative to the information you are displaying. Keep it simple and clear.
- Avoid pie charts, because humans have difficulty perceiving relative angles.
- Pay attention to scale, and make scales consistent.
- Explore several ways to display the data!

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation5.html

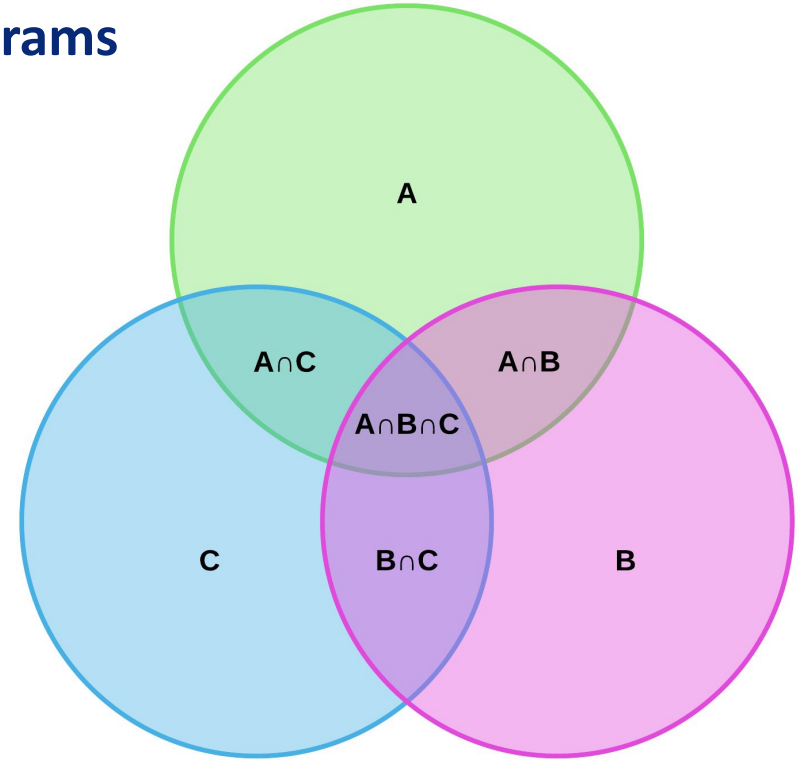
Venn Diagrams

DMI

- ✓ A diagram that shows all possible logical relationships between a finite collection of different sets.
- ✓ Each set is represented by a circle. The circle size sometimes represents the importance of the group but not always.
- ✓ The groups are usually overlapping: the size of the overlap represents the intersection between both groups.
- ✓ A venn diagram makes a really **good work to study the intersection between 2 or 3 sets**. It becomes very hard to read with more groups than that and thus must be avoided.

Venn Diagrams

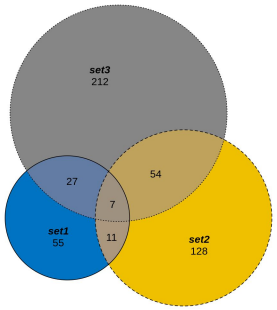
DMI



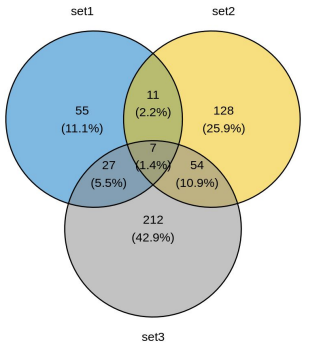
Venn Diagrams in R

DMI

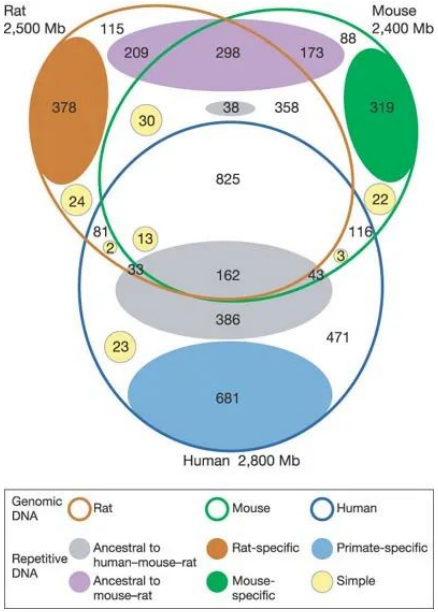
```
set.seed(1)
sets <- list()
sets[["set1"]] <- paste(rep("word_" , 100) , sample(c(1:1000) , 100 , replace=F) , sep="")
sets[["set2"]] <- paste(rep("word_" , 200) , sample(c(1:1000) , 200 , replace=F) , sep="")
sets[["set3"]] <- paste(rep("word_" , 300) , sample(c(1:1000) , 300 , replace=F) , sep="")

library(eulerr)
plot( euler(sets, shape = "circle"),
    quantities = TRUE,
    fill = c("#0073C2FF", "#EFC000FF", "#868686FF"
    lty = 1:3,
      labels = list(font = 4))
```




```
library(ggvenn)
ggvenn(
 sets,
 fill_color = c("#0073C2FF", "#EFC000FF", "#868686FF"),
 stroke_size = 0.5, set_name_size = 4)
```



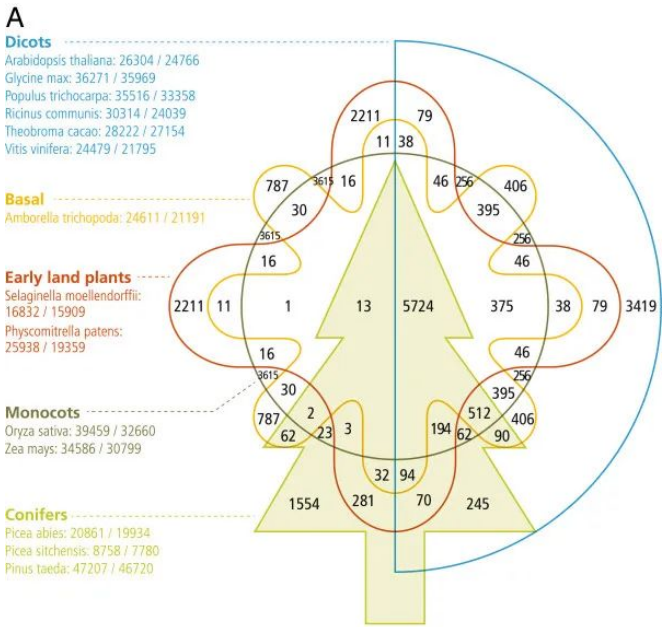
Venn Diagrams



Genome sequence of the Brown Norway rat yields insights into mammalian evolution

https://www.nature.com/articles/nature02426/figures/7

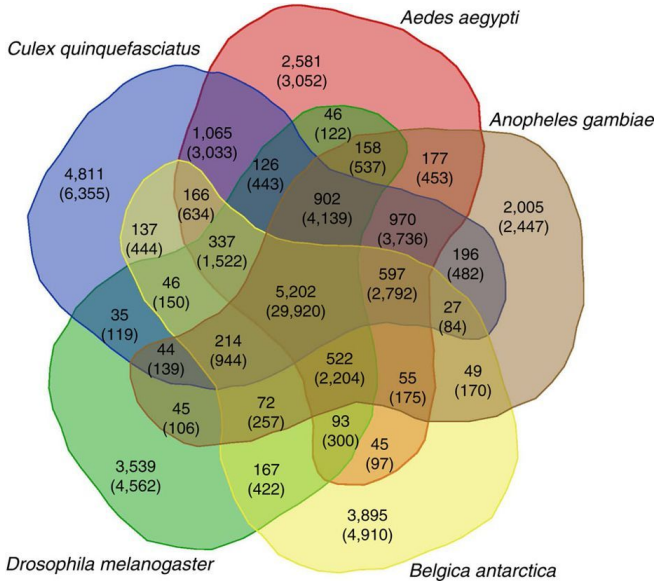
Venn Diagrams



Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r59

Venn Diagrams

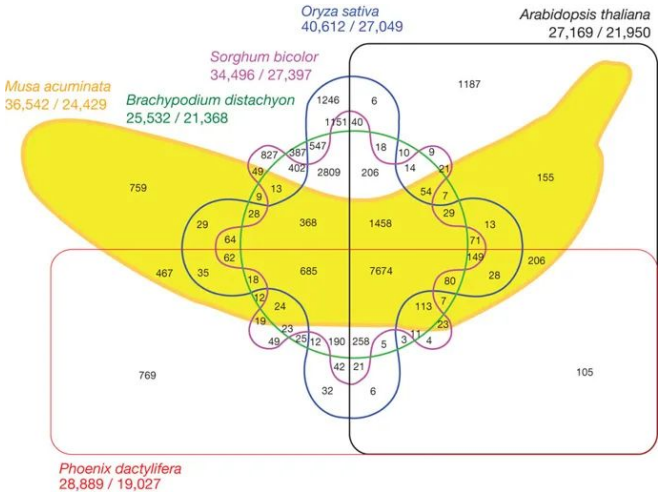


The banana (Musa acuminata) genome and the evolution of monocotyledonous plants

https://www.nature.com/articles/nature11241

http://bioinformatics.psb.ugent.be/webtools/Venn/

Venn Diagrams



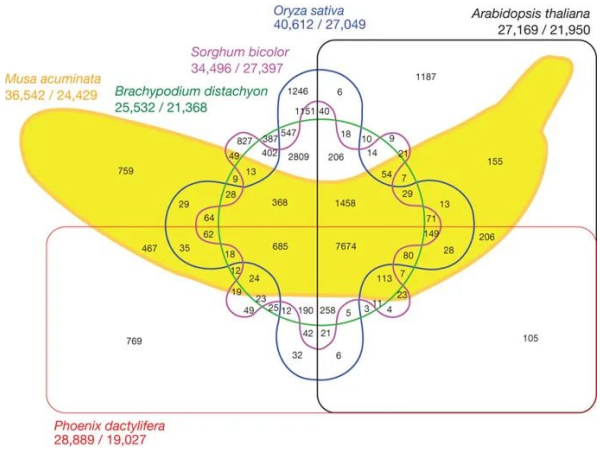
The banana (Musa acuminata) genome and the evolution of monocotyledonous plants

https://www.nature.com/articles/nature11241

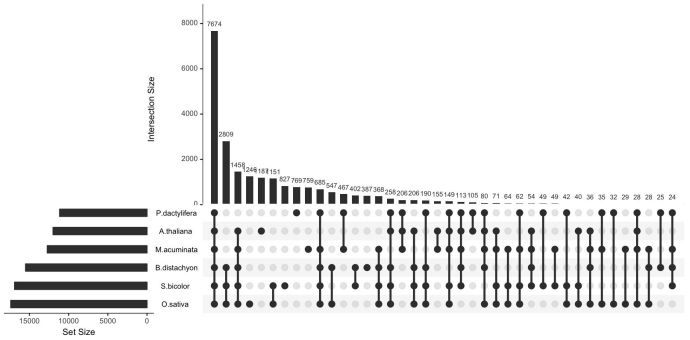
Here is a famous example: a six-set venn diagram published in Nature that shows the relationship between the banana's genome and the genome of five other species.

Venn Diagrams

An alternative to The banana genome Venn diagram



Upset R package
https://github.com/hms-dbmi/UpSetR



Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter

DMI

Design errors	Content errors
Tables that are too large so that it is hard for readers to follow, or too simple so that the information should be included in the text	Inclusion of nonessential data
Failure to use shading and bordering in tables, when both techniques improve readability	Redundancy with text
Incorrect choice of graphical format or scale to portray data	Excessive precision in tables (ie, including too many significant figures)
Use of 3-dimensional graphs when 2 dimensions would suffice	Not self-explanatory (ie, graphic cannot be fully interpreted when isolated from the main text)
Design elements interfere with clarity of graph or figure	Inadequate definition of symbols or abbreviations

https://linkinghub.elsevier.com/retrieve/pii/S0363-5023(11)01653-4

Table or Graph?

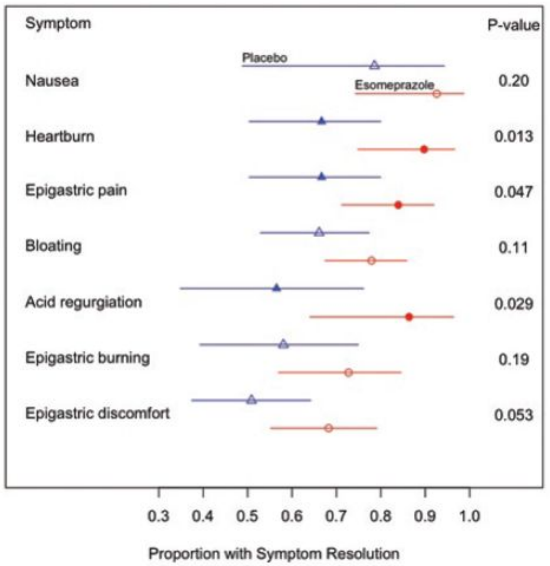
- ✓ Tables are generally best if you want to be able to look up specific information or if the values must be reported precisely.
- ✓ Graphics are best for illustrating trends and making comparisons

Table or Graph?

Proportion of Patients With Resolution of Investigator-Assessed Upper Gastrointestinal Symptoms at 26 wk (Intention-to-Treat Population)

Symptom	Esomeprazole 20 mg n/N (%)	Placebo n/N (%)	*P* Value (CMH)
Epigastric pain	47/56 (83.9)	28/42 (66.7)	0.047
Epigastric burning	32/44 (72.7)	18/31 (58.1)	0.1876
Epigastric discomfort	43/63 (68.3)	29/57 (50.9)	0.0533
Heartburn	35/39 (89.7)	28/42 (66.7)	0.0131
Acid regurgitation	19/22 (86.4)	13/23 (56.5)	0.0290
Nausea	25/27 (92.6)	11/14 (78.6)	0.1988
Bloating	67/86 (77.9)	41/62 (66.1)	0.1126

CMH = Cochran-Mantel-Haenszel χ^2 test (stratified by baseline severity).



Source: Connor JT. Statistical Graphics in AJG: Save the Ink for the Information. Am J of Gastroenterology. 2009; 104:1624-1630.

Principles for Table Display

DMI

- ✓ Sort table rows in a meaningful way
- ✓ Avoid alphabetical listing!
- ✓ Use rates, proportions or ratios in addition (or instead of) totals
- ✓ Multiple time points may be better presented in a Figure
- ✓ Similar data should go down columns
- ✓ Highlight important comparisons
- ✓ Show the source of the data

Principles for Table Display

Type	Incidence	Proportion
Bladder	72,570	5.7%
Breast	232,340	18.2%
Colon	142,820	11.2%
Kidney	59,938	4.7%
Leukemia	48,610	3.8%
Lung	228,190	17.9%
Melanoma	76,690	6.0%
Lymphoma	69,740	5.5%
Pancreas	45,220	3.5%
Prostate	238,590	18.7%
Thyroid	60,220	4.7%

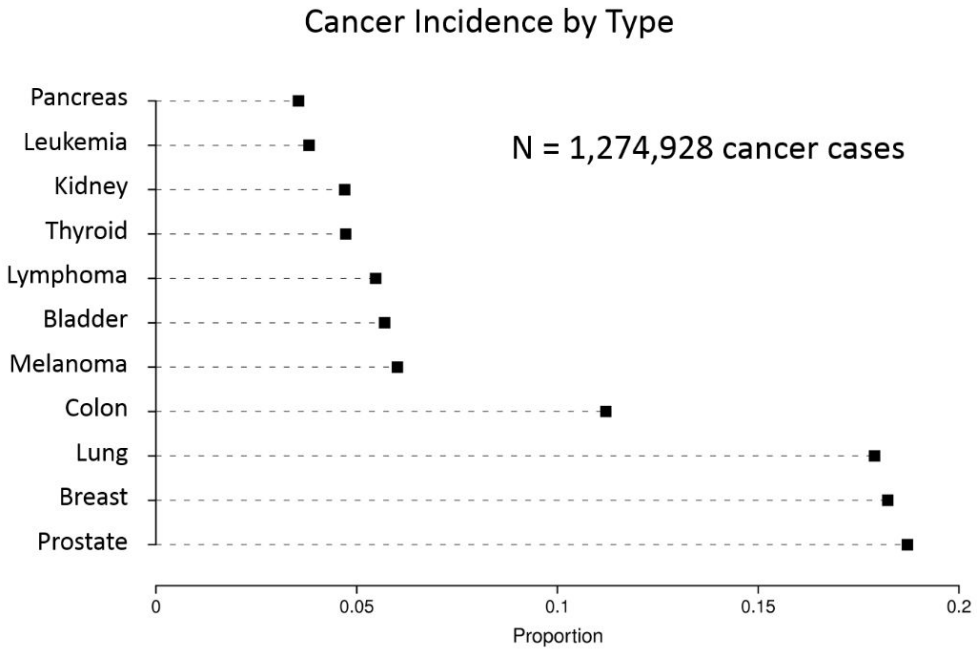
Principles for Table Display

DMI

Type	Incidence	Proportion
Bladder	72,570	5.7%
Breast	232,340	18.2%
Colon	142,820	11.2%
Kidney	59,938	4.7%
Leukemia	48,610	3.8%
Lung	228,190	17.9%
Melanoma	76,690	6.0%
Lymphoma	69,740	5.5%
Pancreas	45,220	3.5%
Prostate	238,590	18.7%
Thyroid	60,220	4.7%

Type	Incidence	Proportion
Prostate	238,590	18.7%
Breast	232,340	18.2%
Lung	228,340	17.9%
Colon	142,820	11.2%
Melanoma	76,690	6.0%
Bladder	72,570	5.7%
Lymphoma	69,740	5.5%
Thyroid	60,220	4.7%
Kidney	59,938	4.7%
Leukemia	48,610	3.8%
Pancreas	45,220	3.5%

Principles for Table Display



Data from http://www.cancer.gov/cancertopics/types/commoncancers

DMI

Visualization of Biomedical Data: selected resources for biomedical scientists

Resource	Description	URL
Discovery		
Plotly	Online tool for fast data visualization	https://plot.ly/create/
Tableau	For interactive visualizations, including web based	http://tableau.com
Matplotlib	For tailored visualizations of data sets in Python	http://matplotlib.org
ggplot2	For tailored visualizations of complex data sets in R	http://ggplot2.org
D3.js	For tailored, interactive web-based visualizations	https://d3js.org
Communication		
Photoshop	For editing imaging data	http://adobe.com/photoshop
GIMP	Free, open-source alternative to Photoshop	http://www.gimp.org
Illustrator	For creating and editing vector graphics	http://adobe.com/illustrator
Inkscape	Free, open-source alternative to Illustrator	http://inkscape.org
Utilities		
Color Brewer	Web tool for selecting contrasting color maps	http://colorbrewer2.org
Adobe Color	Web tool for designing sets of colors	http://color.adobe.com
Paletton	Web tool for designing sets of colors	http://paletton.com
General Resources		
BioVis	Computer science publications on biological visualizations	http://biovis.net
Clarafic	Training guides for biomedical visualization tools	http://clarafi.com
InformationisBeautiful	Showcase of charts and infographics for a wide variety of data	http://bit.ly/Info_Beauty
Visual Complexity	Catalog of tailored visualizations for complex data	http://visualcomplexity.com
VIZBI	Collected videos and posters on tailored biological visualizations	http://vizbi.org

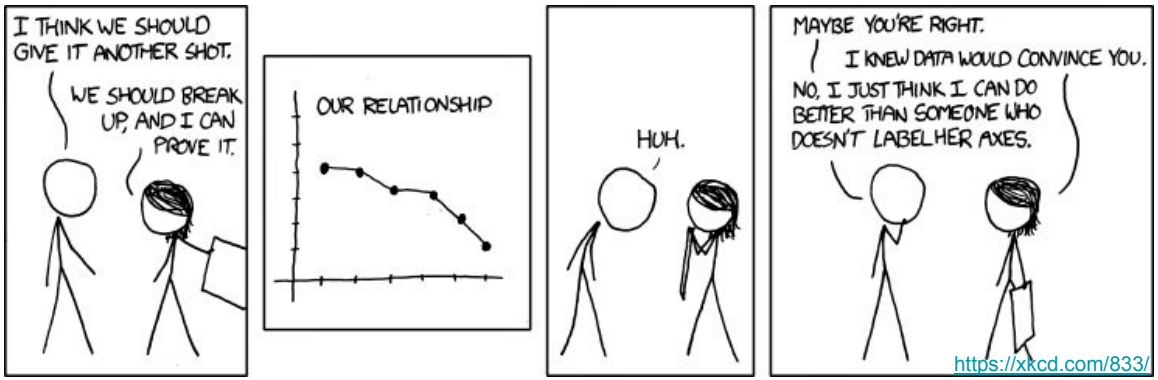
10.1146/annurev-biodatasci-080917-013424

Summary

DMI

- Put the things to be compared next to each other
- Use color to set things apart, but consider color blind folks
- Use position rather than angle or area to represent quantities
- Align things vertically to ease comparisons
- Use common axis limits to ease comparisons
- Use labels rather than legends
- Sort on meaningful variables (not alphabetically)
- Must 0 be included in the axis limits?
- Consider taking logs and/or differences

And remember to label your axes



Figures Checklist

- Does each figure communicate an important piece of information or address a question of interest?
- Do all your figures include plain language axis labels?
- Is the font size large enough to read?
- Does every figure have a detailed caption that explains all axes, legends, and trends in the figure?

Hands-on session

Go to this page and accept the assignment (if you have not done it yet!)

Remember to take into account these principles and suggestions for all the assignments in the course!

Deadline to submit the assignment: 15/03/2026

Bibliography

Book: The Visual Display Of Quantitative Information by Edward Tufte

Visualization of Biomedical Data (2018)

https://doi.org/10.1146/annurev-biodatasci-080917-013424

Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter (2012) https://doi.org/10.1016/j.jhsa.2011.12.041

Ten simple rules to create biological network figures for communication (2019)

https://doi.org/10.1371/journal.pcbi.1007244