

Data Visualization

Applied Data Analysis and Visualization I

Department of Methodology and Statistics
Javier Garcia-Bernardo

Today

What

Introduction

Data visualization

Model fit and cross validation

Linear regression for data science

Classification

Interactive Data Visualization

...

When

Week 1

Week 2

Week 3

Week 4

Week 5

Week 6

...

Main points for today

- Part 1: **Why** do we use visualizations
- Part 2: How to make **good data visualizations**
 - **A grammar of graphics** (Wickham) and `ggplot2`
 - Perception: Visual channels and type of plots
 - Design: Principles of design
 - Storytelling: Use pre-attentive attributes to guide the reader
- Part 3: Conclusions

PART 1: Why data visualizations

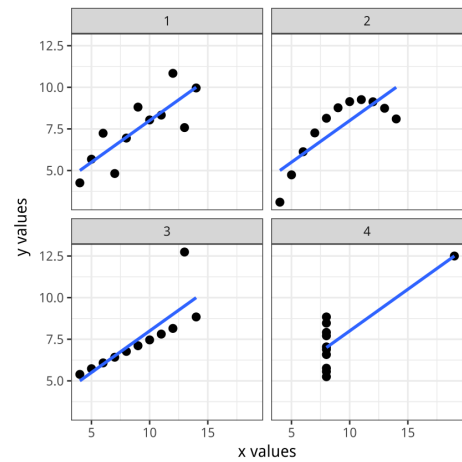
Data visualization

- If done correctly: Extremely efficient way of processing and remembering data
 - Reduces cognitive load: Mental effort needed to process and understand information
- Two goals of visualization
 - Explore the data:
 - Understand distributions, outliers and relationships
 - Communicate a message:
 - Main focus of this lecture
- Often the only part of the analysis that the reader ever sees.

Example: Anscombe's quartet

Anscombe's quartet (*Anscombe, 1973; Chatterjee & Firat, 2007*):

- Four datasets
- Visualization of relationship between 2 quantities (x and y)
- The mean and standard deviation of each x and y variables (e.g., means) are almost identical
- Relationship quantified by the correlation coefficient equals 0.81 for all pairs



Data visualization

- Reducing **cognitive load** makes the audience:
 - More willing to **read** your analysis
 - More likely to **understand** the data/results
 - More prone to **accept** the results
 - More likely to **remember** them

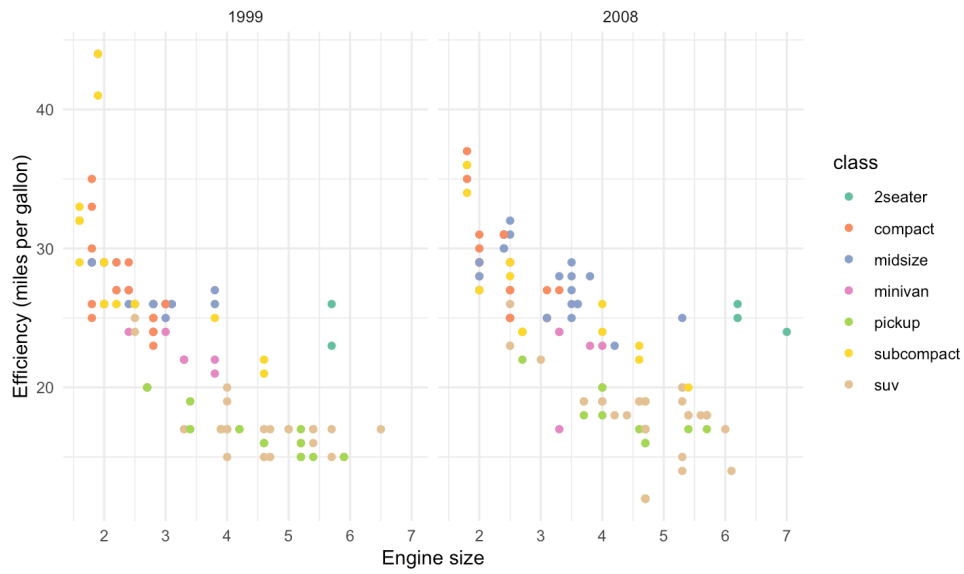
- How do we make sure that the graphs we make transfer:
 - The right part of the data, and;
 - with the least effort possible? (i.e., to **minimize cognitive load**)

Part 2: How to make a good data visualization?

Main points for today

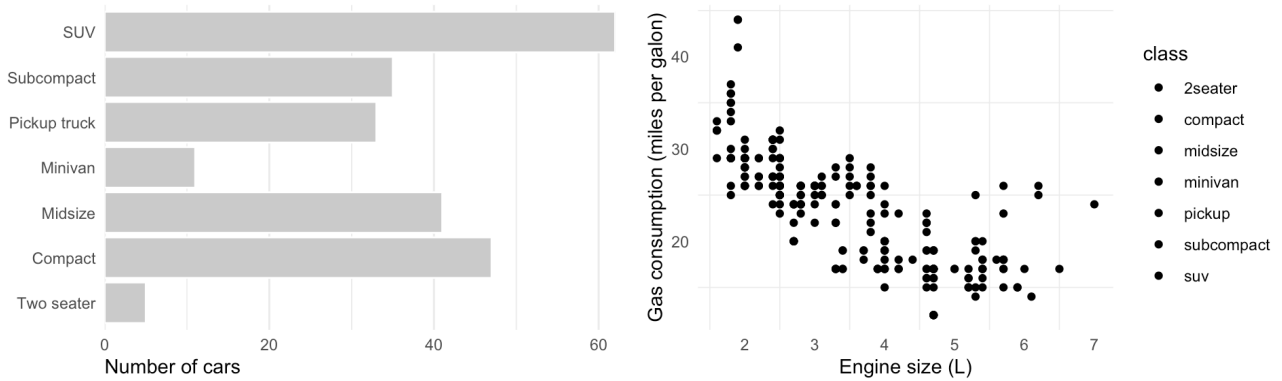
2.1 What are the main elements of a graph?

- We will talk about how to decompose this graph into “pieces” (e.g. labels, dots, etc)



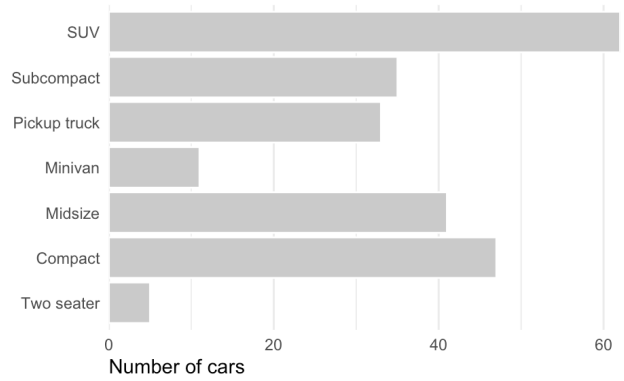
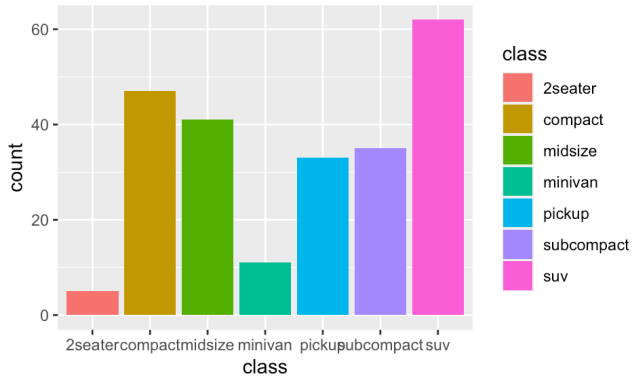
2.2 What type of plot should you use?

- For example we often use **barplots** to compare quantities, **scatterplots** to show relationships, and **lineplots** to track a variable over time.



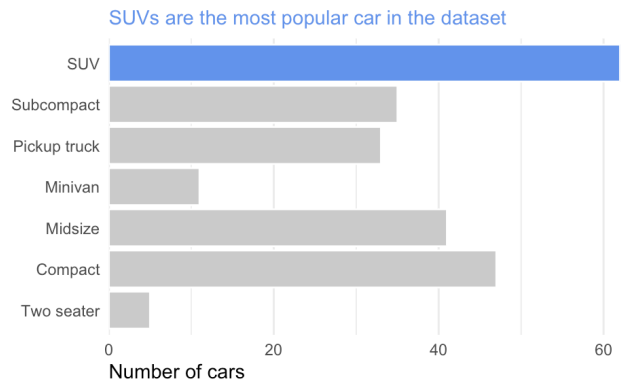
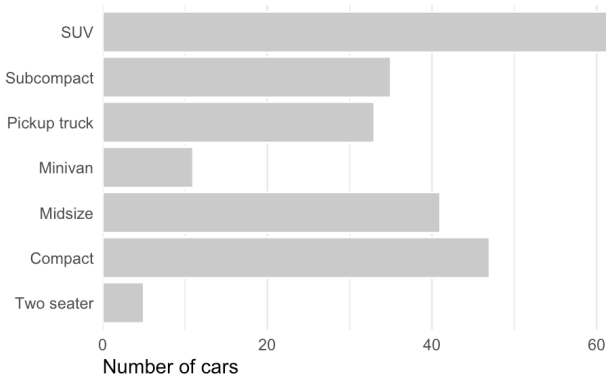
2.3 How can we make a plot look more professional?

- Write down four ways the plot in the right has improved the plot in the left



2.4 How to guide the reader?

- Compare how you read the figure in the right and in the left



2.0 Starting point

- Every visualization should have one main message (and only one)
- The visualization is designed to communicate that message efficiently
- It helps to write down the message (e.g. on the title of the plot)

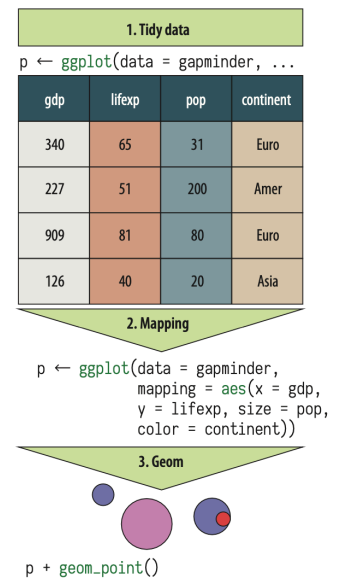
2.1 A grammar of graphics

Grammar of graphics ([Wickham's version](#))

A tool to break up the task of making a graph into a series of subtasks

In R, grammar of graphics is implemented in `ggplot()`, a function in the `ggplot2` package.

- Elements of a graph:
 - The data: `ggplot(data = gapminder)`
 - Aesthetic mappings (position, shape, color, ...) – map variables to influence visual channels: `mapping = aes(x = gdp, y = pop)`
 - Geometric objects (points, lines, bars, ...) – use those mappings: `+ geom_point()`
 - Labels (titles, caption, axes labels): `+ labs(x = "GDP", y = "Population")`



Source: Healy (2019). [Original paper](#)

Grammar of graphics ([Wickham's version](#))

- Additionally, can apply:
 - Scales (linear, logarithmic, ...)
 - Facets (small multiples / subplots)
 - Statistical transformation (identity, binning, unique, jitter, ...)
 - Coordinate system (Cartesian, polar, parallel, ...)

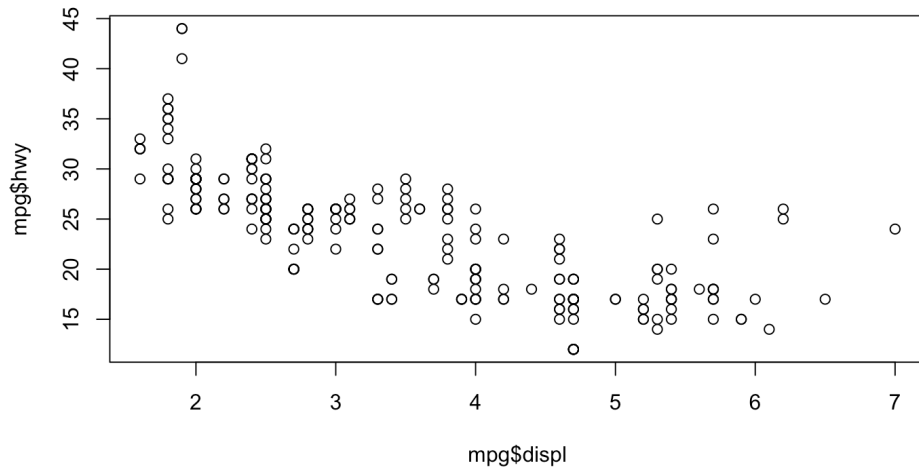
Example data set: mpg (about cars)

```
## # A tibble: 234 × 11
##   manufacturer model      displ  year  cyl trans  drv      cty   hwy fl      class
##   <chr>          <chr>    <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
## 1 audi          a4         1.8  1999     4 auto... f        18    29 p    comp...
## 2 audi          a4         1.8  1999     4 manu... f        21    29 p    comp...
## 3 audi          a4         2     2008     4 manu... f        20    31 p    comp...
## 4 audi          a4         2     2008     4 auto... f        21    30 p    comp...
## 5 audi          a4         2.8  1999     6 auto... f        16    26 p    comp...
## 6 audi          a4         2.8  1999     6 manu... f        18    26 p    comp...
## 7 audi          a4         3.1  2008     6 auto... f        18    27 p    comp...
## 8 audi          a4 quattro  1.8  1999     4 manu... 4        18    26 p    comp...
## 9 audi          a4 quattro  1.8  1999     4 auto... 4        16    25 p    comp...
## 10 audi         a4 quattro  2     2008     4 manu... 4        20    28 p    comp...
## # i 224 more rows
```

- displ: engine displacement, in litres
- hwy: highway miles per gallon
- class: "type" of car

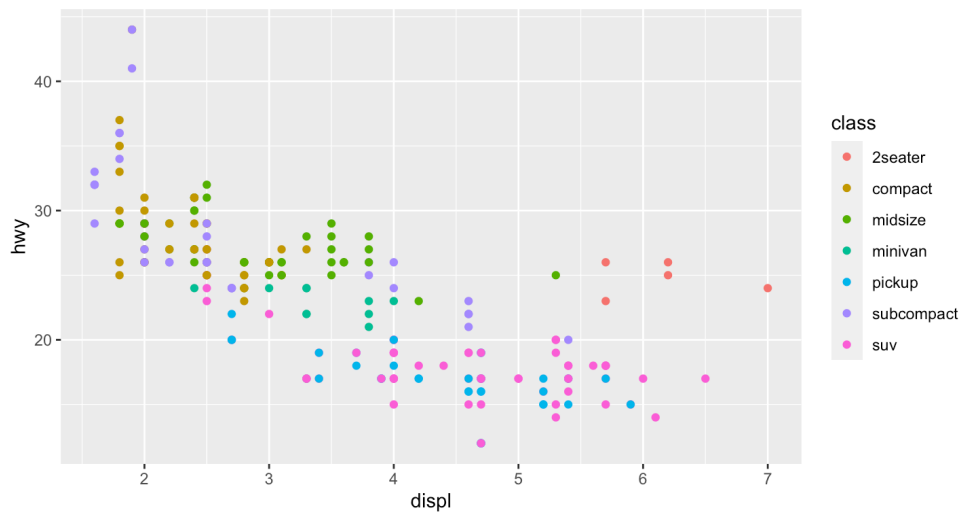
Example data set: mpg - basic plot

```
plot(x = mpg$displ, y = mpg$hwy)
```

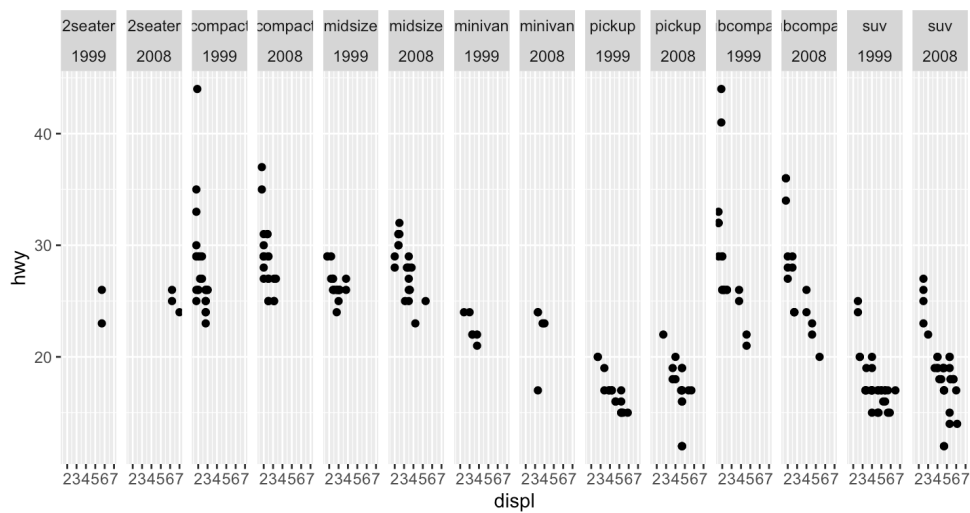


Example data set: mpg - basic ggplot

```
ggplot(data = mpg,  
       mapping = aes(x = displ,  
                     y = hwy,  
                     color = class)) +  
geom_point()
```

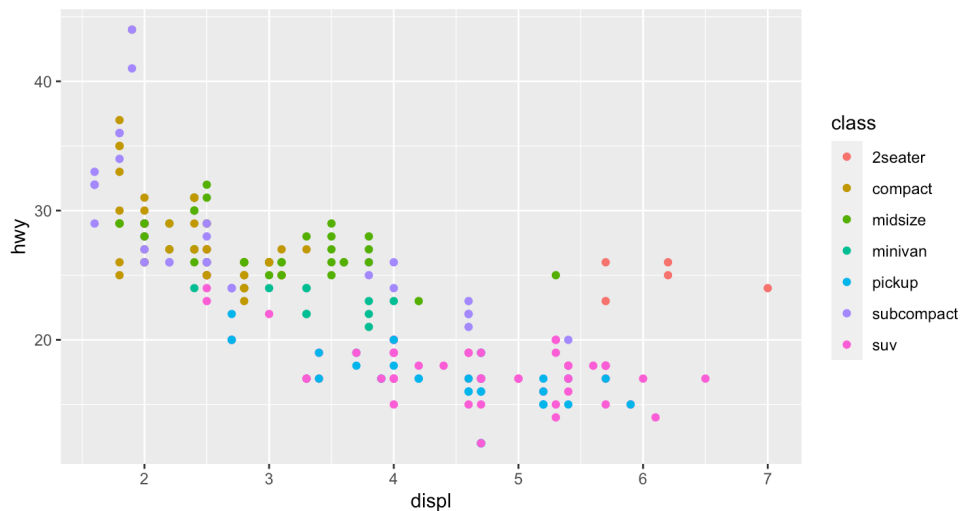


```
ggplot(data = mpg,  
       mapping = aes(x = displ,  
                     y = hwy)) +  
geom_point() + facet_grid(~ class + factor(year))
```

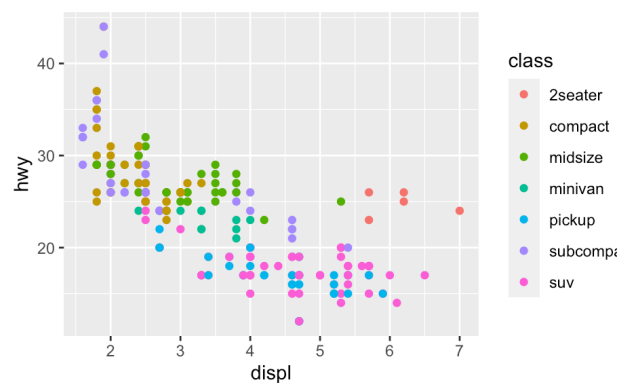


Example data set: mpg - basic ggplot

```
ggplot(data = mpg,  
       mapping = aes(x = displ,  
                     y = hwy,  
                     color = class)) +  
geom_point()
```

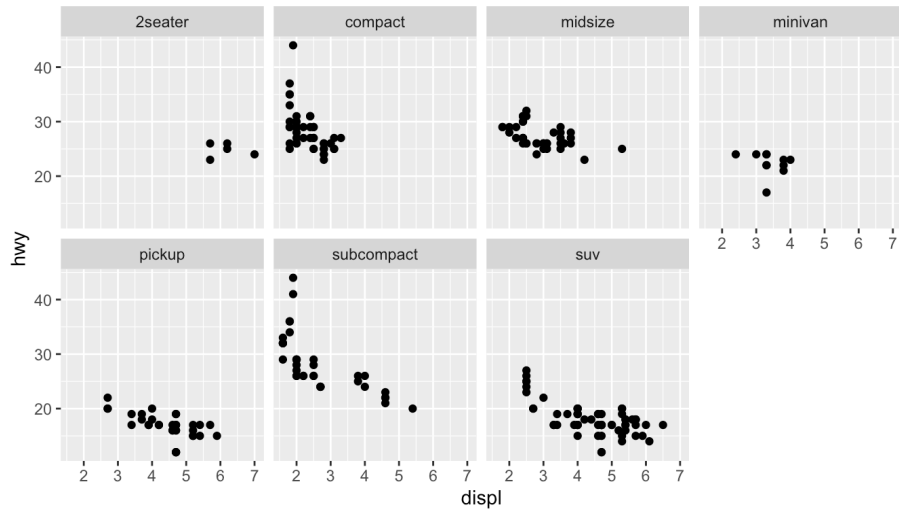


- **Aesthetics** (mapping data to channels):
 - x-position mapped to engine size (`displ`)
 - y-position mapped to fuel efficiency (`hwy`)
 - color mapped to car type (`class`)
- **Geometric objects:** points
- **Transformation:** none (identity)
- **Scales:** continuous, cartesian coordinates
- **No facets**



Facets

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +  
  geom_point() +  
  facet_wrap(~ class, nrow = 2)
```

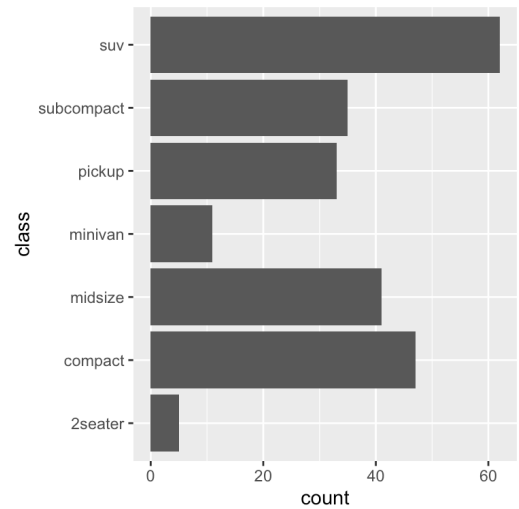


Statistical transformations

```
ggplot(data = mpg, mapping = aes(y = class)) +  
  geom_bar()
```

Total number of cars per car type:

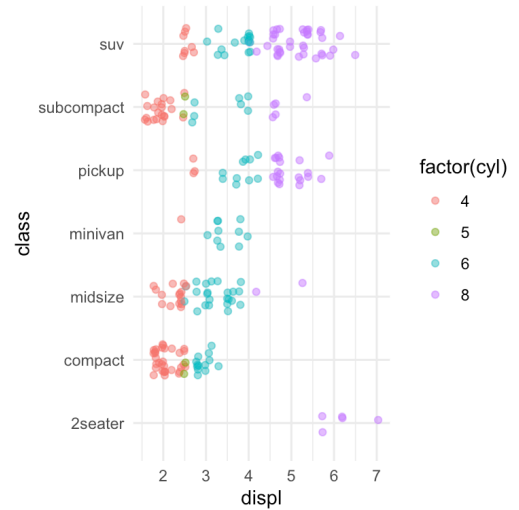
- Aesthetics:
 - y-position mapped to car type (**class**)
- Geometric objects: bars
- Transformations:
 - **geom_bar()** transforms the data, visualizing the **count** of each group. Length of the bars made proportional to the number of cars in each group.



Example data set: mpg - basic ggplot

```
ggplot(data = mpg, mapping = aes(x = displ, y = class, color = factor(cyl))) +  
  geom_point(alpha = 0.5, position = "jitter") +  
  theme_minimal()
```

- Aesthetics (mapping data to channels):
 - x-position mapped to ?
 - y-position mapped to ?
 - color mapped to ?
 - Geometric objects: ?
 - Transformation: ?
- Why do we need factor(cyl)?



Which aesthetics are there?

Colour, fill, alpha (opacity/transparency)

Linetype, linewidth, size, shape, angle

Position: x, y, xmin, xmax, ymin, ymax, xend, yend

Label, fontface, family

<https://ggplot2.tidyverse.org/reference/index.html#section-aesthetics>

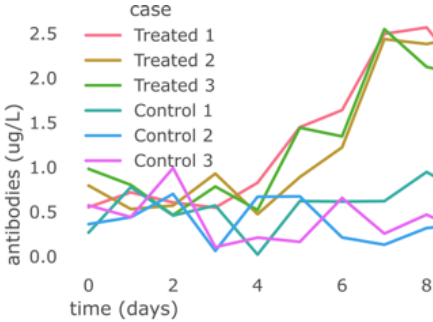
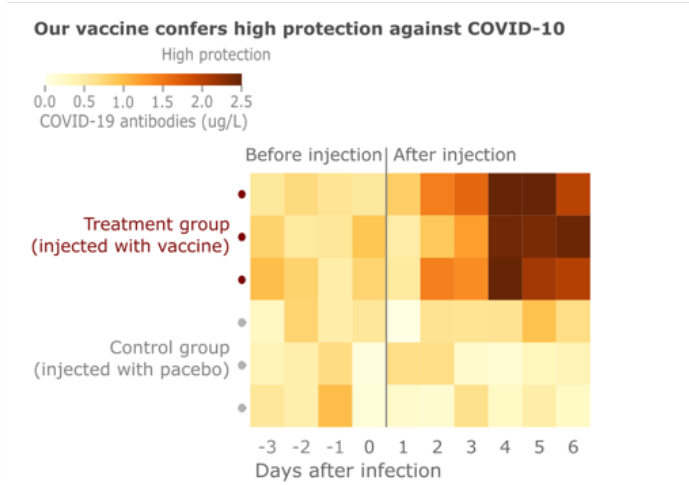
Which geoms are there?

Lines, bars, bins, boxplots, contours, densities, dotplots, error bars, hexagons, polygons, histograms, jittered points, ranges, maps, paths, points, ribbons, areas, rugmarks, segments, curves, smooths, spokes, labels, text, rasters, rectangles, tiles, violins

<https://ggplot2.tidyverse.org/reference/index.html#section-geoms>

Exercise

- Choose a plot (left or right). Write down 1—3 things you like, 1—3 things you would change



Check the effectiveness of different channels
(aesthetics): tinyurl.com/uu-dataviz

10 minute break



Topics

Before the break

- Part 1: Why do we visualize data
- Part 2: How to make a good data visualization
 - 2.1 A grammar of graphics

Now

- Part 2: How to make a good data visualization
 - 2.2 Visual channels and type of plots
 - 2.3 Principles of design
 - 2.4 Guiding the reader using pre-attentive attributes
- Part 3: Conclusion

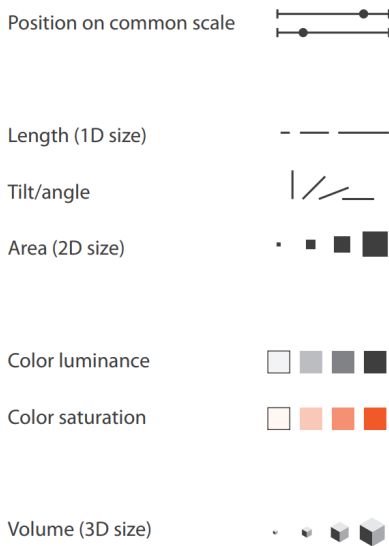
2.2 Visual channels & type of plots

Some visual channels are more effective

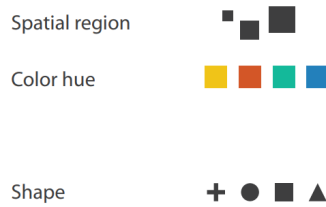
Now we know to make a plot with `ggplot()`. But how do we know which visual channels (aesthetics) and type of plots (geometry) to use?

Channels: Expressiveness Types and Effectiveness Ranks

➔ **Magnitude Channels: Ordered Attributes**

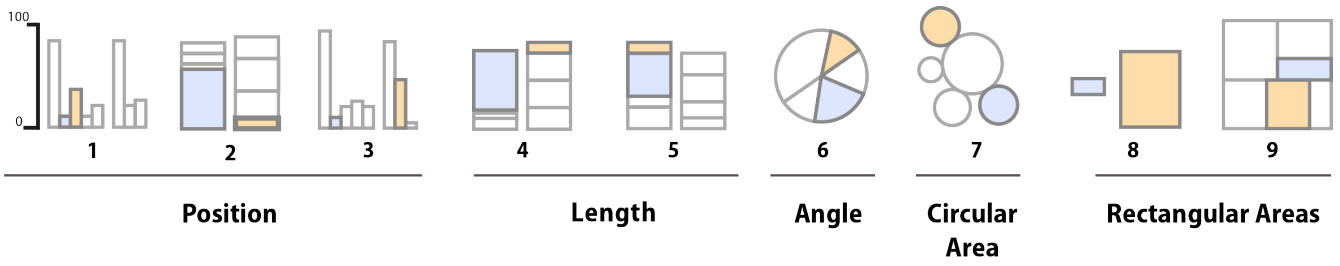


➔ **Identity Channels: Categorical Attributes**



Visualization Analysis and Design (Tamara Munzner)

For quantitative variables



Let's see how you did

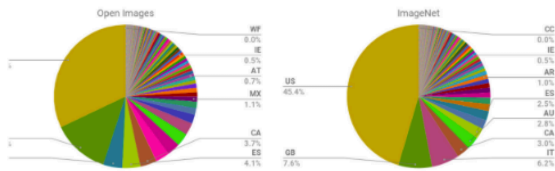


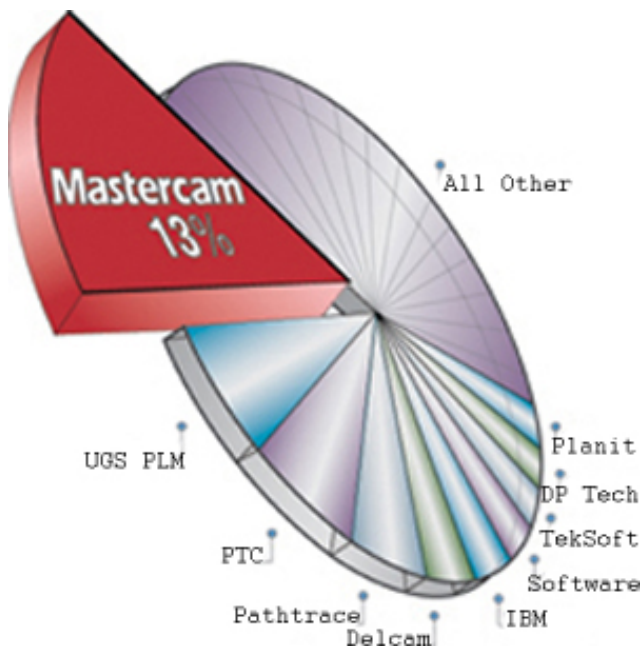
Fig. 3. Fraction of each country, represented by their two-letter ISO codes, in Open Images and ImageNet image datasets. In both datasets, US and Great Britain represent the top locations (from Reference [138], © Shreya Shankar).



Fig. 4. Geographic distribution of countries in the Open Images dataset. In their sample, almost one third of the data was US-based, and 60% of the data was from the six most represented countries across North America and Europe (from Reference [138], © Shreya Shankar).

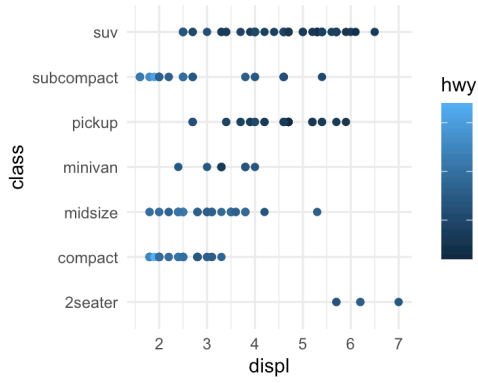
- Main message of both plots: Show **accurately** the main contributors to the dataset (in number of images)
- Think about what channels are more efficient to transfer quantitative data.
- Would you use a:
 - Barplot (one bar per country, countries can be aggregated)
 - Stacked barplots (one bar, in chunks, per country)
 - Pie chart
 - Treemap (nested rectangular areas)
 - Map + color

Do not use 3D. Do not use angles.

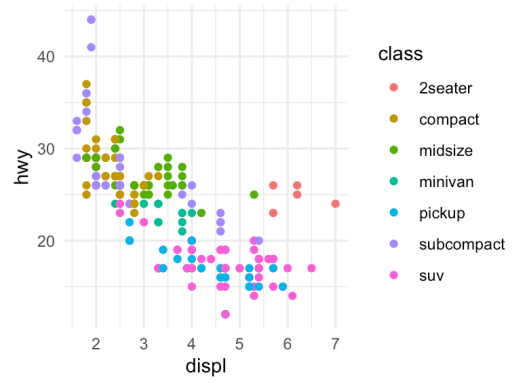


- Message you want to convey: Compact and subcompact cars are very efficient
- What are the most important variables? (displ: engine size; hwy: miles per gallon; class: type of car)

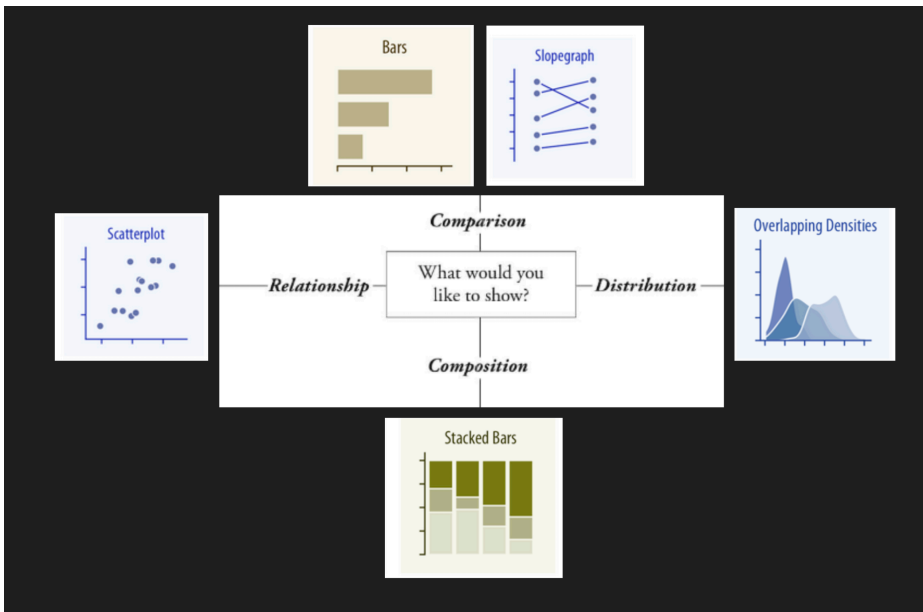
```
ggplot(data = mpg,
       mapping = aes(y = class,
                     x = displ,
                     color = hwy)) +
  geom_point() + theme_minimal()
```



```
ggplot(data = mpg,
       mapping = aes(y = hwy,
                     x = displ,
                     color = class)) +
  geom_point() + theme_minimal()
```

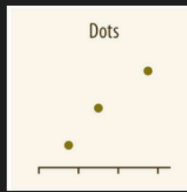
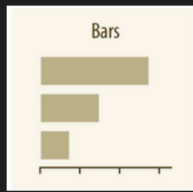


Channels and type of graph



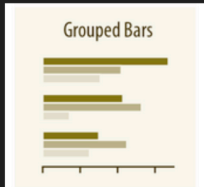
Fundamentals of Data Visualization by Claus O. Wilke

AMOUNTS AND PROPORTIONS

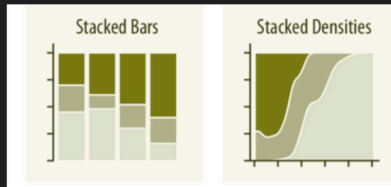


Required with log-scales

Trends



Differences within row



Proportions over x

Fundamentals of Data Visualization by Claus O. Wilke

DISTRIBUTIONS

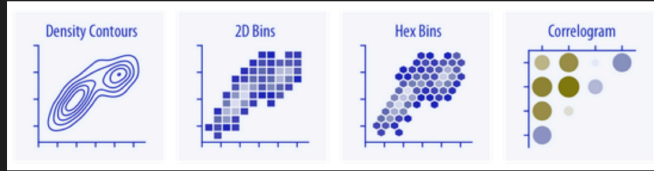


Fundamentals of Data Visualization by Claus O. Wilke

RELATIONSHIPS



Too many points?



Time series?

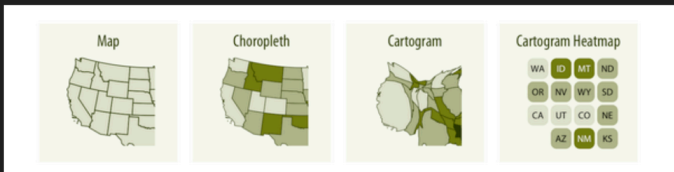


Fundamentals of Data Visualization by Claus O. Wilke

GEOGRAPHICAL DATA

- ▶ Color is key (more on this later)
- ▶ Combine with a barplot or bubbles if the values are important


Do you *actually* need a map?





Fundamentals of Data Visualization by Claus O. Wilke


2.3 Principles of design

CONTRAST

COLOR 

TOPE/VALUE 

SIZE/SHAPE 

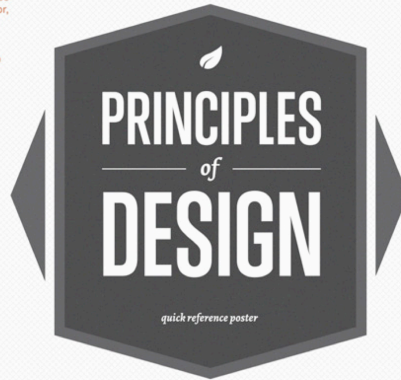
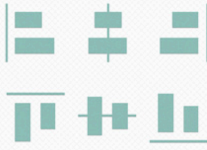
DIRECTION 

Unique elements in a design should stand apart from one another. One way to do this is to use contrast. Good contrast in a design - which can be achieved using elements like color, tone, size, and more - allows the viewer's eye to flow naturally.

To the left, you can see 4 ways to create contrast in your design.

ALIGNMENT

Proper alignment in a design means that every element in it is visually connected to another element. Alignment allows for cohesiveness; nothing feels out of place or disconnected when alignment has been handled well.

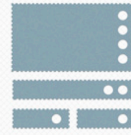


REPETITION

Repetition breeds cohesiveness in a design. Once a design pattern has been established - for example, a dotted border or a specific typographic styling - repeat this pattern to establish consistency.

The short version?

Establish a style for each element in a design and use it on similar elements.



PROXIMITY

Proximity allows for visual unity in a design. If two elements are related to each other, they should be placed in close proximity to one another. Doing so minimizes visual clutter, emphasizes organization, and increases viewer comprehension.

Imagine how ridiculous it would be if the proximity icons on this graphic were located on the other side of this document.

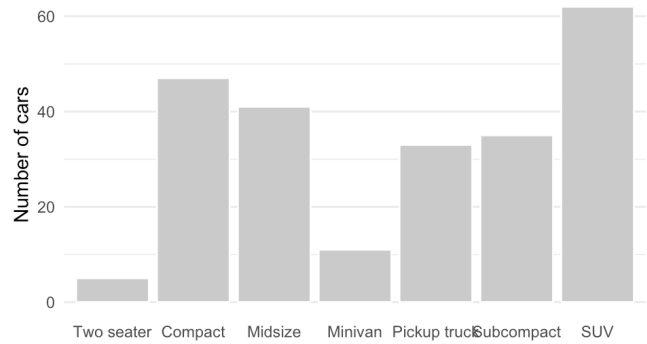
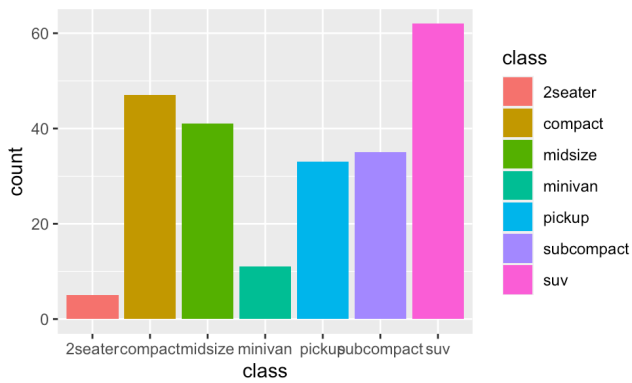


a handy *paperleaf* resource

Source: Paper leaf

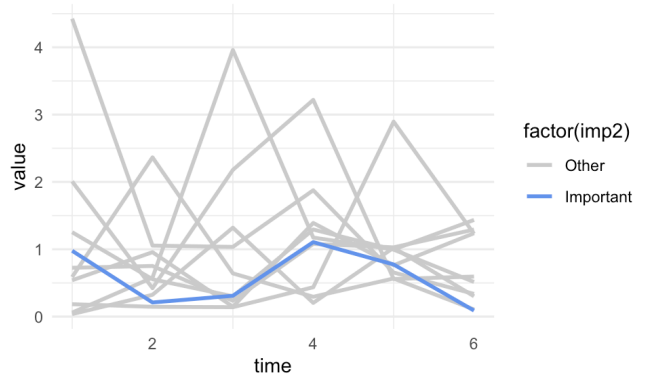
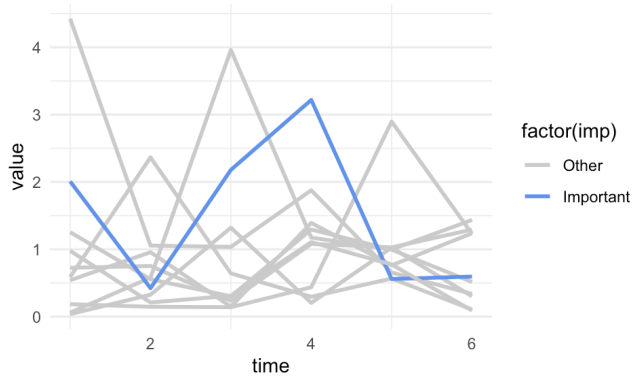
2.3.1 CONTRAST

- Idea: Unique elements should stand apart from each other
- How: Increase contrast and eliminate clutter (maximize data-to-ink ratio)



2.3.2 REPETITION

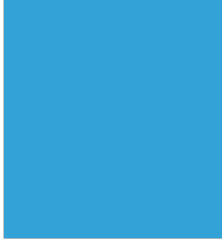
- Repetition creates cohesion
- Make sure the same information is always presented in a coherent way (e.g. same color)



2.3.3 ALIGNMENT

- Proper alignment: every element is visually connected to another elements
- Increases cohesion
- Humans like “strong” lines

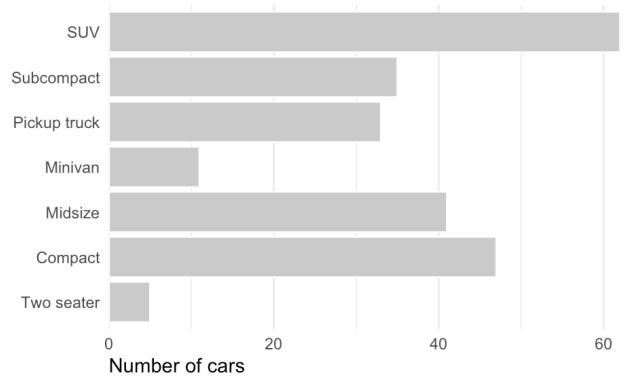
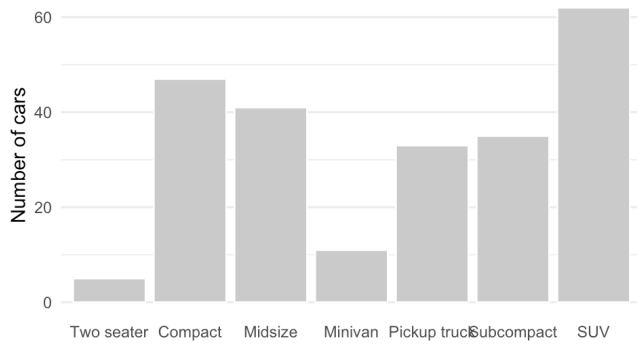
**LOOKS
GOOD**



**DOES NOT
LOOK SO GOOD**

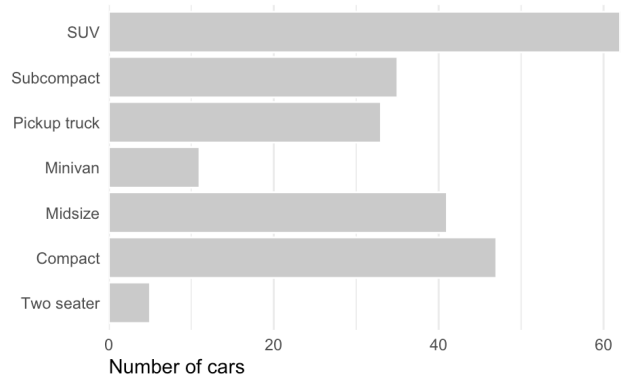
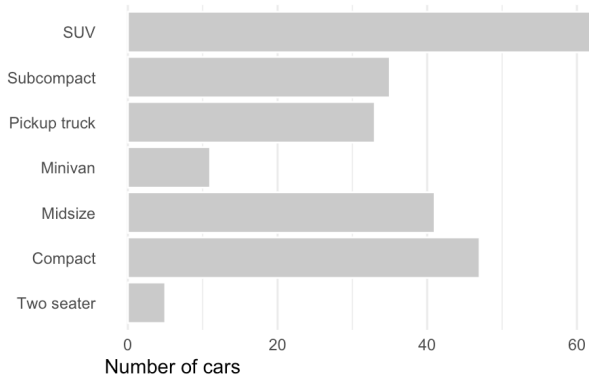


2.3.3 ALIGNMENT

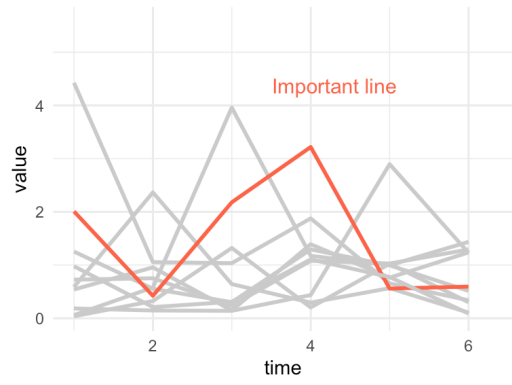
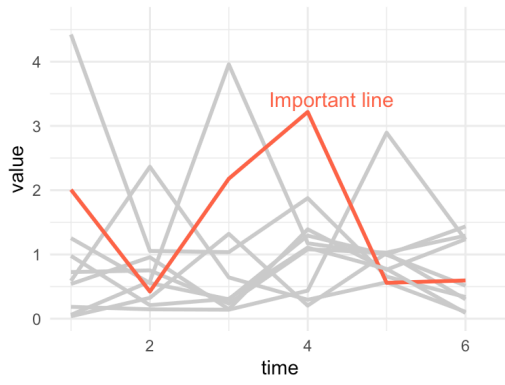


2.3.4 PROXIMITY

- Proximity reduces clutter and organizes the space

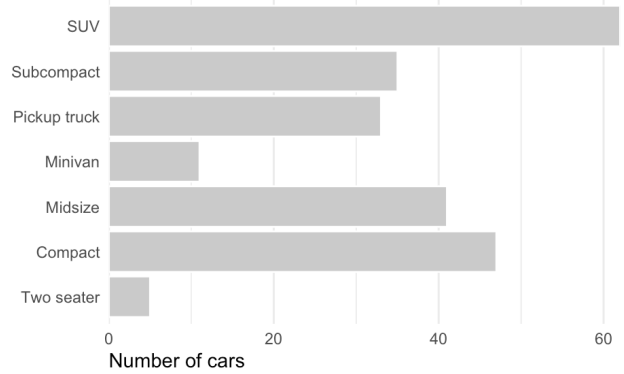
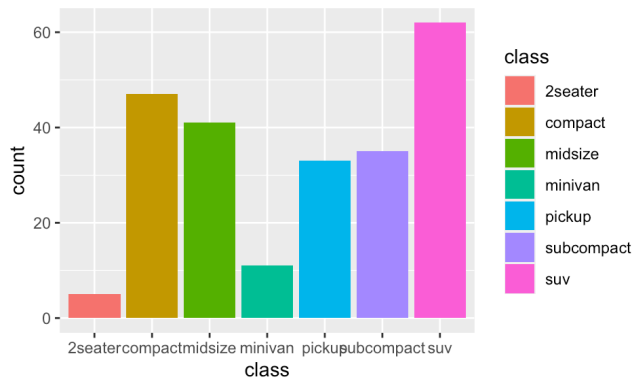


2.3.4 PROXIMITY



How can we make a plot look more professional?

- Write down four ways (CRAP) the plot in the right has improved the plot in the left



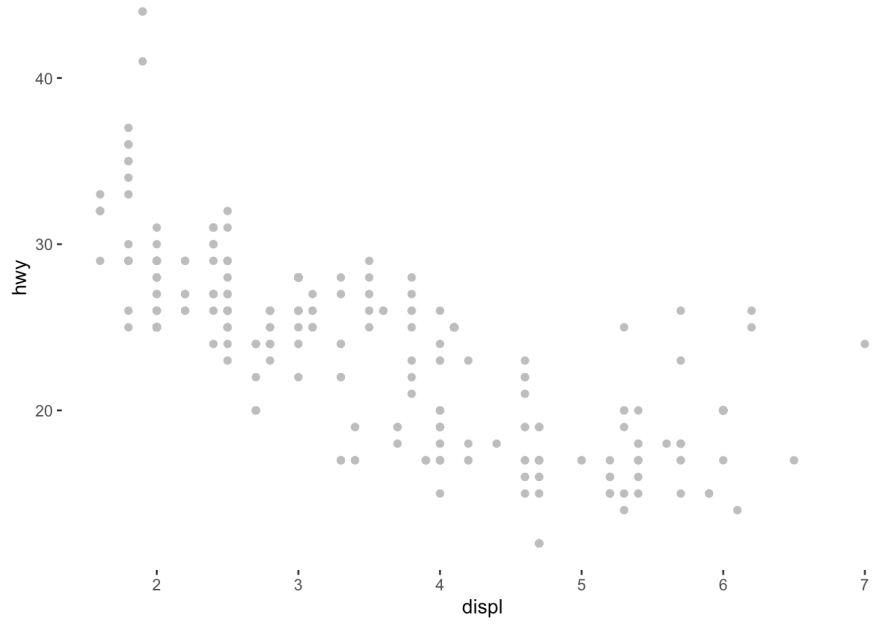
Practical advise about design

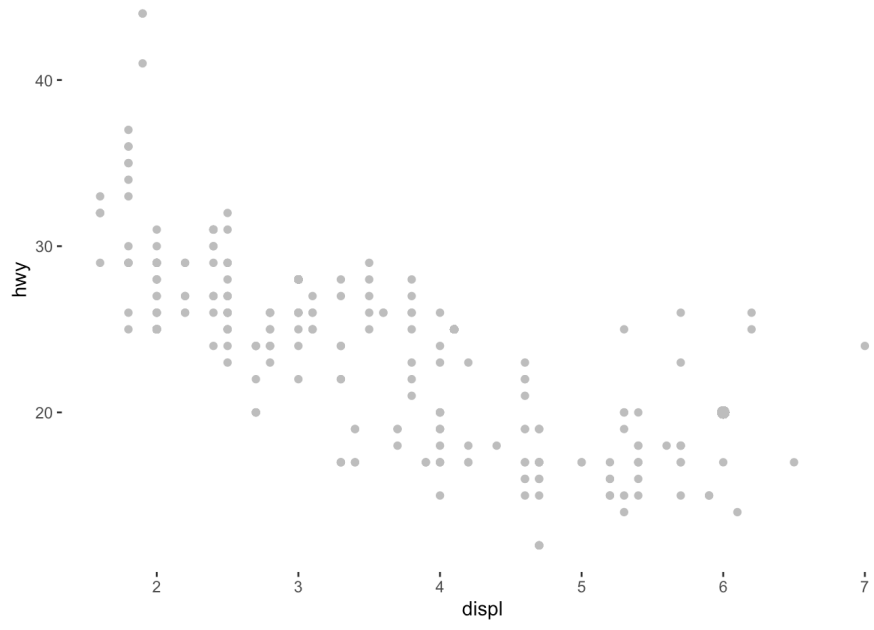
- Reduce cognitive load:
 - Removing unnecessary clutter
 - More professional/aesthetically pleasant
- Contrast:
 - Eliminate unnecessary lines (all frames, use gray grid lines, etc)
 - Don't use a gray background
 - White space is your friend (allows for "breathing")
 - Enlarge the labels
 - Use vector graphics (svg/pdf/eps) to avoid blurry figures -> Edit them in Illustrator or Inkscape
- Repetition: Be consistent across figures
- Alignment: Make sure you align subplots/labels
- Proximity: When possible, label data directly (instead of using legends)

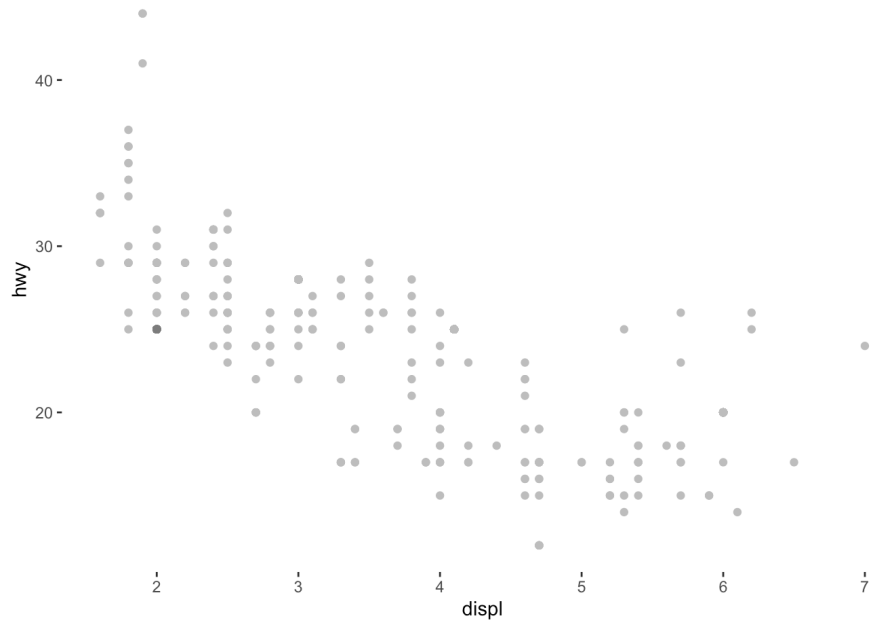
2.4 Pre-attentive attributes to guide the reader

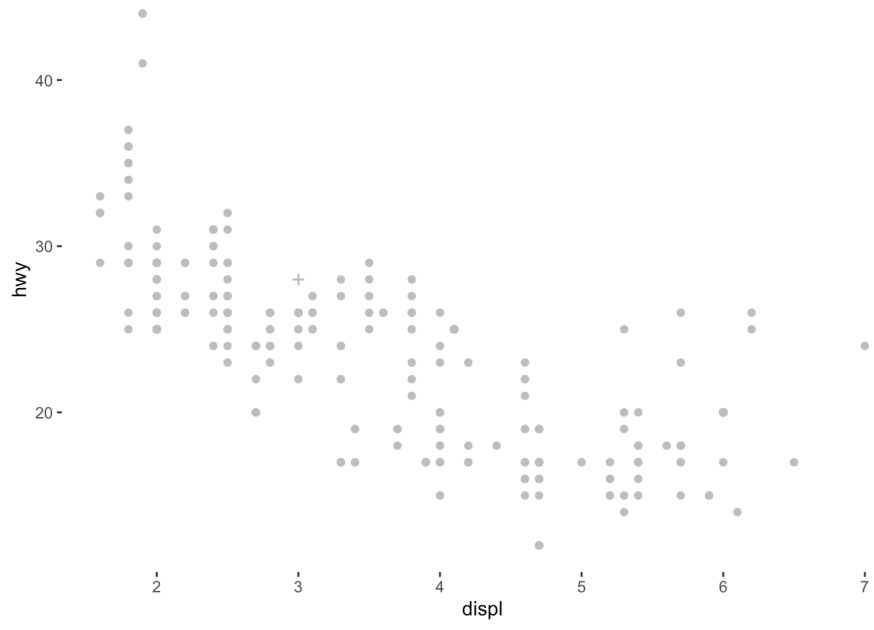
Guiding the reader

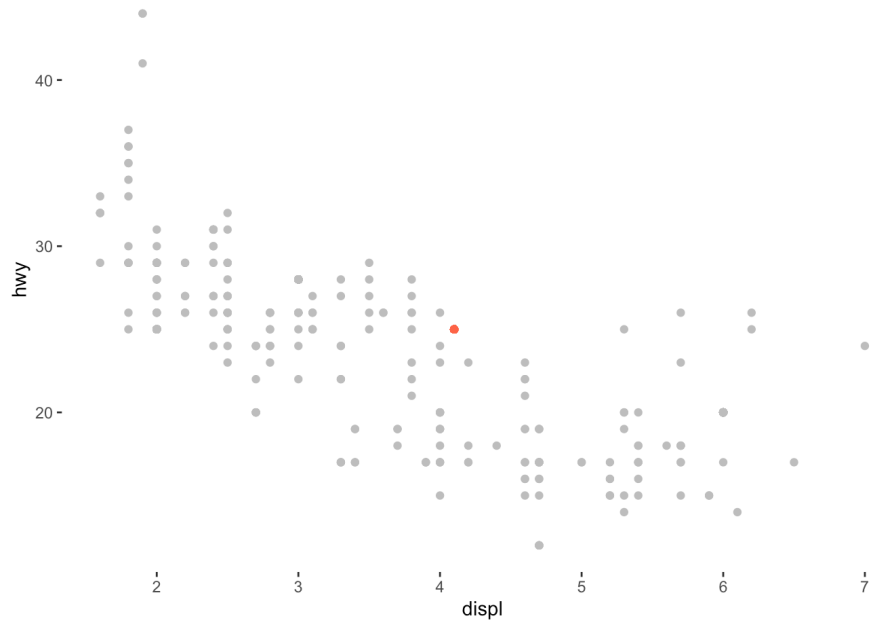
- Why: Help the person interpret the plot (reduce cognitive load, make it enjoyable)
- How: Use pre-attentive attributes (elements that “pop” without searching for them)





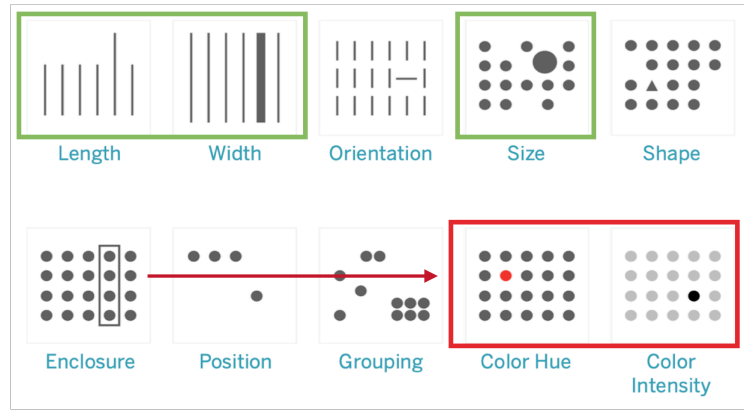






What do we focus on:

- ▶ Large objects
- ▶ Bright objects
- ▶ Contrasting objects



Source: Storytelling with data

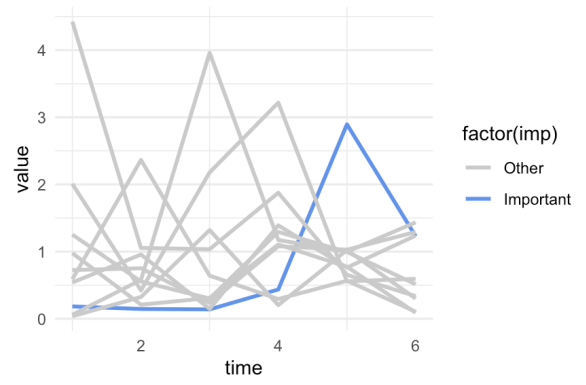
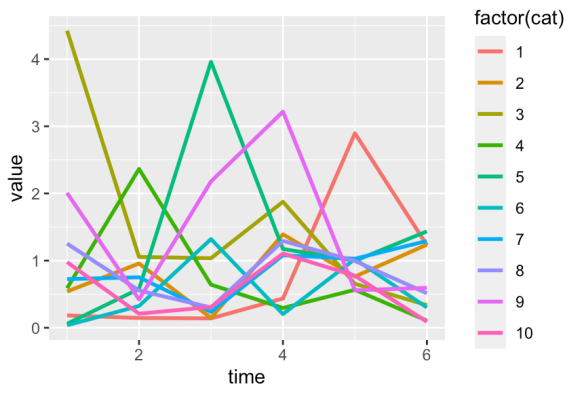
Color makes ice cream taste sweeter
veggies taste fresher,
and coffee taste richer

— Ellen Lupton

COLOR

- Two uses of color:
 - To encode data
 - To guide attention
 - Color is the most useful pre-attentive attribute
 - It also allows for consistency across figures
- Color affect emotion and this is culture-dependent. Some responses are nearly universal
 - Warm colors -> alive/alert
 - Blue colors -> calming/focus
- More information: <https://blog.datawrapper.de/which-color-scale-to-use-in-data-vis/>
<https://davidmathlogic.com/colorblind/#%23D81B60-%231E88E5-%23FFC107-%23004D40>

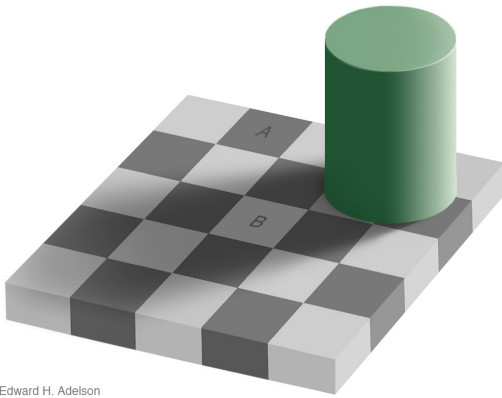
Color to guide attention



Color to encode data

- In addition of highlighting, colours can be used to:
 - Represent categories (no more than 4-5 colors)
 - Represent quantitative values:
 - Only if necessary (i.e. you need to use the x and y axis for more important variables)
 - Not accurate (still okay if you only want to show trends)

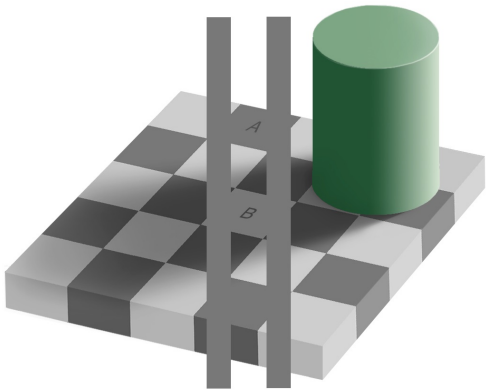
Colors are perceived relatively



Edward H. Adelson

- The same shade of gray will be perceived very differently depending on whether it is against a darker background or a lighter one.

Colors are perceived relatively



- The same shade of gray will be perceived very differently depending on whether it is against a darker background or a lighter one.
- In addition, we are better at distinguishing darker shades than we are at distinguishing lighter ones.

Colors are perceived relatively

- Colors are a mix of
 - *luminance* or *lightness*: (relative) brightness,
 - *hue*: amount of red, green, blue, and
 - *chroma* or *saturation*: intensity or vividness of the color
- To represent quantitative values: Want mappings from data to color that are perceptually uniform
- Default palettes available in R (e.g. `ggplot2` or in the package `Rcolorbrewer`) are perceptually uniform.

The Original Default "Rainbow (Jet)" Colormap



The Recent Default "Viridis" Colormap



Source: <https://nightingaledvs.com/color-in-a-perceptual-uniform-way/>

Use a color palette that fits your message



GUIDE THE READER

- Think about the main message of your visualization
- Think about the way the reader will interact with the plot
 - Use color and other pre-attentive attributes to draw attention to the important parts
 - We read plots in a Z-shaped flow: top-left to top-right to bottom-left to bottom-right

You probably read this 1st

You probably read this 2nd

You probably read this 3th

You probably read this 4th

You probably read this 1st

You probably read this 2nd



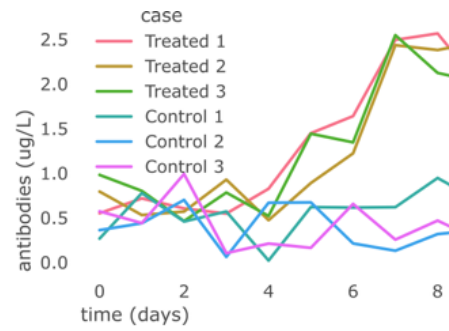
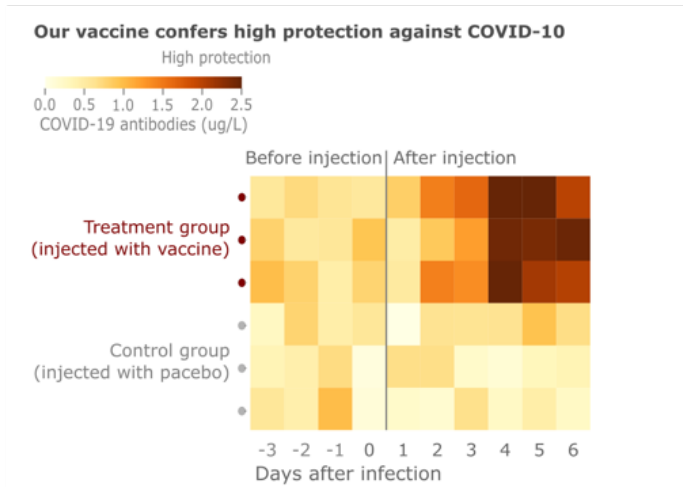
You probably read this 3th

You probably read this 4th



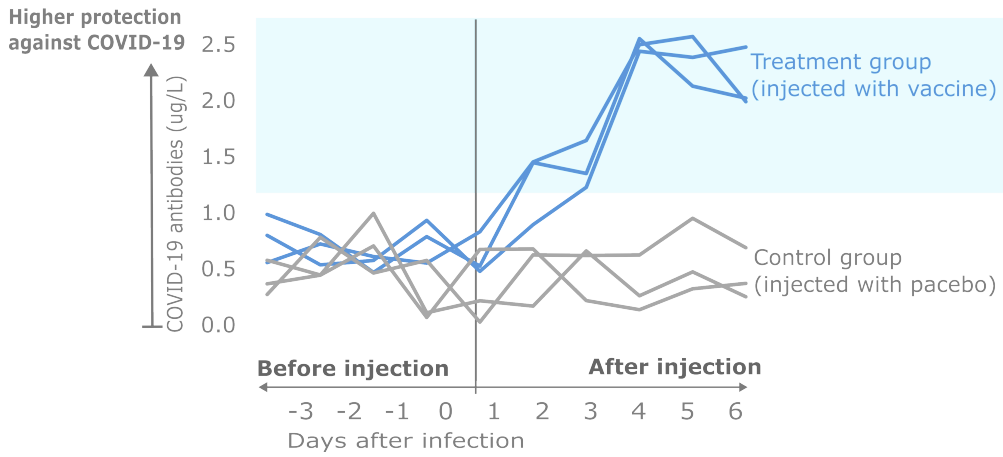
Which plot is better?

- Type of plot?
- Design principles?



Which plot is better?

Our vaccine confers high protection against COVID-10



Part 3: Conclusion & Practical guide

Conclusions

- **Data visualization**
 - Efficient and effective to show data → Reduce cognitive load
 - Sticking to basic principles helps (e.g., by using `ggplot`).
- **Why do we want to reduce cognitive load**
 - More willing to read your report
 - More likely to understand the data/results
 - More willing to accept the results
 - More likely to remember them (and you)

Bad graphs and how to make them better:

- **Substantive:** problems due to the data being presented
 - Understand the data and check its quality
 - Use tidy data to minimize the risk of errors
 - Think about the main message and the intended audience/media
- **Perceptual:** graph is confusing or misleading because of how people perceive and process what they are looking at
 - Use the “grammar of graphics”
 - Use the right type of plot (map important variables to efficient visual channels)
 - Guide the reader: Focus attention with pre-attentive attributes
 - Guide the reader: Add labels and annotations
- **Aesthetic:** tacky, tasteless, inconsistent, ugly plots
 - Use the CRAP principles of design
 - Save the figure as PDF (or EPS)
 - Do minor edits in Canvas, Illustrator or Inkscape

Conclusions and Good practice

- When constructing a graphic, consider the following:
 - What is the main message and the intended audience?
 - What are: aesthetics, geom, scale, facets, transformation, coordinate system?
 - Are your most important variables mapped to unbiased channels (length/position)?
 - Can you remove clutter (increase data/ink)?
 - Are you guiding the reader via preattentive attributes to show the main message?
 - Do I visualize the data in a honest way (e.g., am I including the context in a time series)?

Next class

How to judge if your model is good: **model fit** and **cross-validation**.

Have a nice day!