

Supervised Learning and Visualization: Exploratory Data Analysis

Gerko Vink
Erik-Jan van Kesteren
Peter van der Heijden

Department of Methodology and Statistics



Universiteit Utrecht

Utrecht Applied Data Science



Data Scientist patient zero

Inventor of:

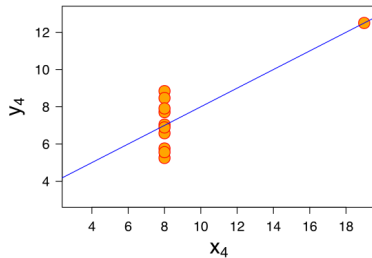
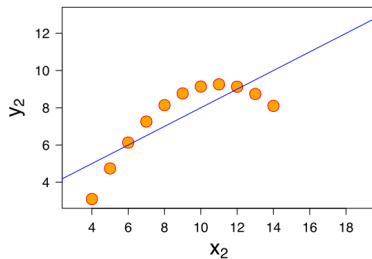
- The boxplot
- The term “exploratory data analysis”
- The Fast Fourier Transform
- “Tukey’s test”
- The word “bit”
- So, so much more (Wikipedia)

Today: visualization principles, applicable to EDA

Some data visualization principles

Data visualization

- For exploration, data analysis ←
- For communication
- For entertainment



C. J. F. B. M.

- We'd like to make, not just any kind of picture or graph, but one that transfers some part of the data to our brain
- How do we make sure that the graphs we make transfer
 - ① The right part of the data, and;
 - ② As much of it as possible?

This is where the “**grammar of graphics**” comes in.

Goal is to **specify how data map to picture**, so the correct type and largest amount possible is transferred

- Aesthetics (position, shape, color, ...)
- Geometric objects (points, lines, bars, ...)
- Scales (continuous, discrete, ...)
- Facets (small multiples)

- Statistical transformation (identity, binning, median, ...)
- Coordinate system (Cartesian, polar, parallel, ...)

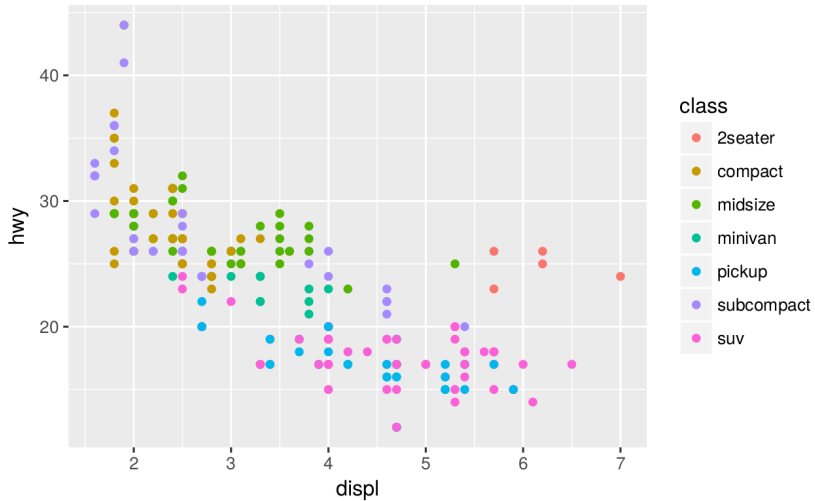
Grammar of graphics (Wickham version)

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

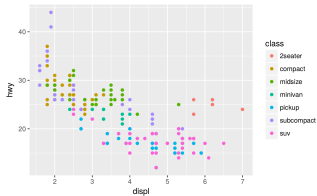
```
#> # A tibble: 234 × 11
```

```
#>      manufacturer model displ  year   cyl      trans      drv      cty   hwy  fl
#>      <chr>    <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>
#> 1      audi      a4    1.8  1999     4    auto(l5)      f    18    29    p
#> 2      audi      a4    1.8  1999     4 manual(m5)      f    21    29    p
#> 3      audi      a4    2.0  2008     4 manual(m6)      f    20    31    p
#> 4      audi      a4    2.0  2008     4    auto(av)      f    21    30    p
#> 5      audi      a4    2.8  1999     6    auto(l5)      f    16    26    p
#> 6      audi      a4    2.8  1999     6 manual(m5)      f    18    26    p
#> # ... with 228 more rows, and 1 more variables: class <chr>
```

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



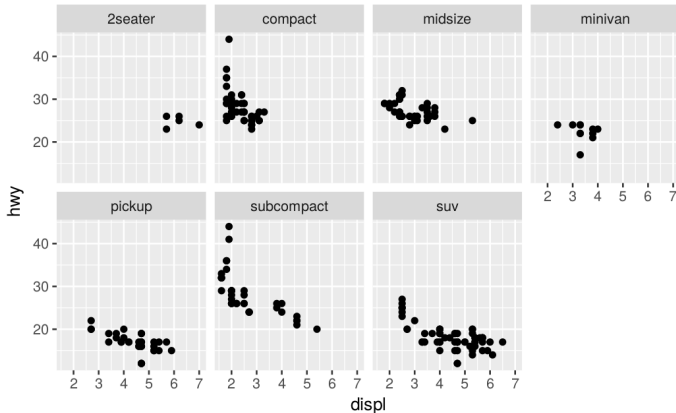
```
ggplot(data = mpg) +
```



- Aesthetics:
 - x-position mapped to *engine displacement*
 - y-position mapped to *highway miles per gallon*
 - color mapped to car type
- Geometric objects: points
- Transformation: identity
- Scales: continuous, cartesian coordinates
- No facets

Facets

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



Transformation (stats)

1. `geom_bar()` begins with the **diamonds** data set

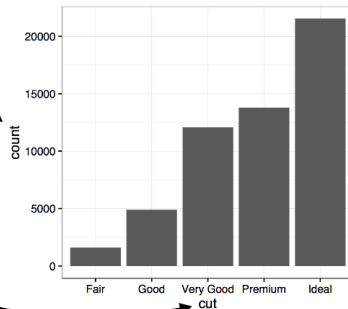
carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
...

`stat_count()`

2. `geom_bar()` transforms the data with the "count" stat, which returns a data set of cut values and counts.

cut	count	prop
Fair	1610	1
Good	4906	1
Very Good	12082	1
Premium	13791	1
Ideal	21551	1

3. `geom_bar()` uses the transformed data to build the plot. cut is mapped to the x axis, count is mapped to the y axis.



What should I choose?

LES VARIABLES DE L'IMAGE

		POINTS			LIGNES			ZONES	
XY 2 DIMENSIONS DU PLAN									
Z	TAILLE								
	VALEUR								

LES VARIABLES DE SÉPARATION DES IMAGES

Z	GRAIN								
	COULEUR								
	ORIENTATION								
	FORME								

Jacques Bertin (1967) *Sémiologie graphique*

Color: hue-saturation-brightness (HSB)

Hue Changes



Saturation Changes



Brightness Changes



Mackinlay's ranking of encodings

Quantitative

Position
Length
Angle
Slope
Area
Volume
Density
Color saturation
Color hue
Texture
Connection
Containment
Shape

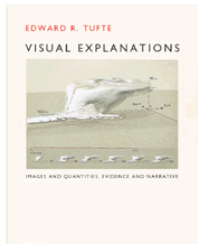
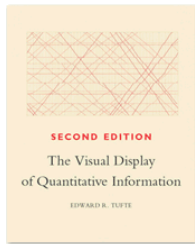
Ordinal

Position
Density
Color saturation
Color hue
Texture
Connection
Containment
Length
Angle
Slope
Area
Volume
Shape

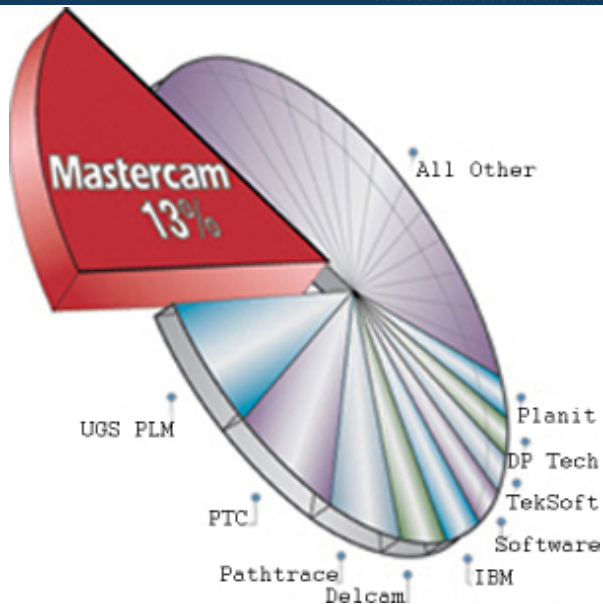
Nominal

Position
Color hue
Texture
Connection
Containment
Density
Color saturation
Shape
Length
Angle
Slope
Area
Volume

Some (distilled) principles from Tufte

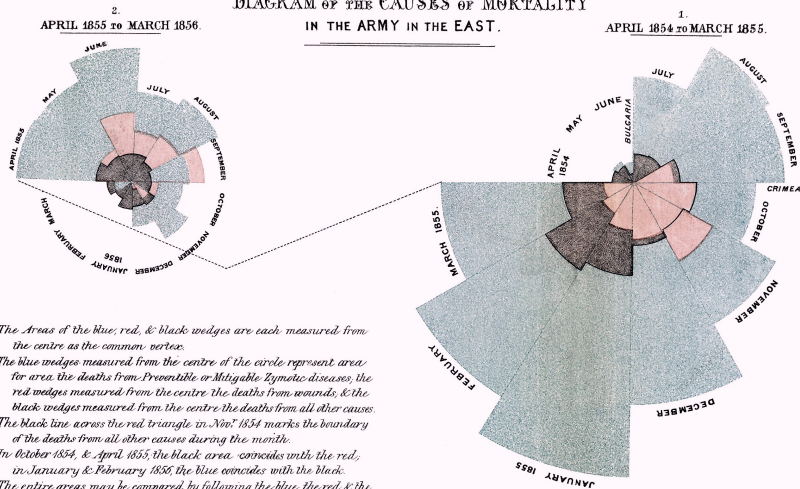


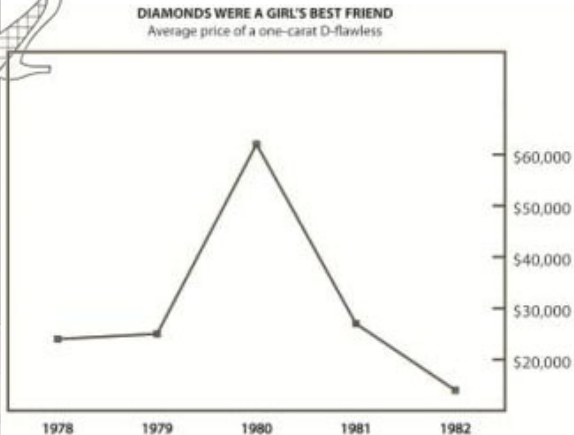
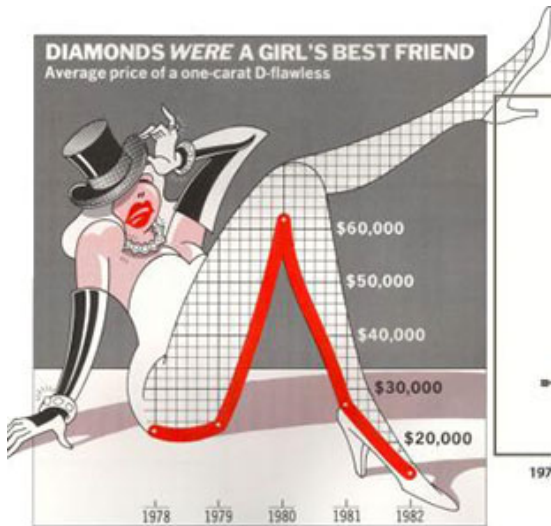
- Ask how data maps to perception
- Ask which comparisons you want, guide eye to those
- Maximize **data-to-ink ratio**
- Present more data (but without losing interpretability)
- (Remember narrative)




Nightingale Rose / Coxcomb chart

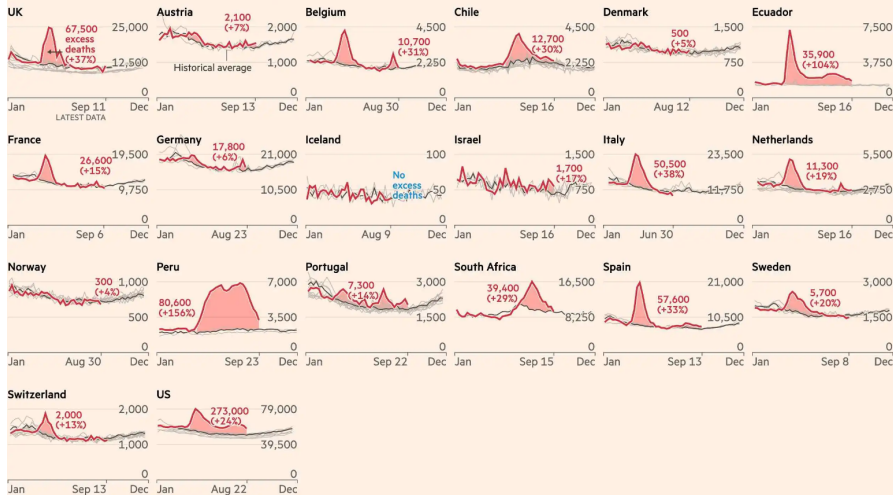
DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.



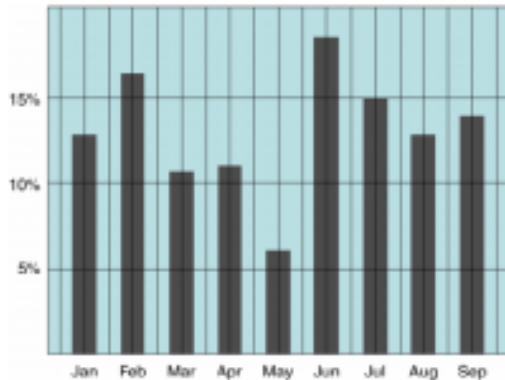


Death rates have climbed far above historical averages in many countries that have faced Covid-19 outbreaks

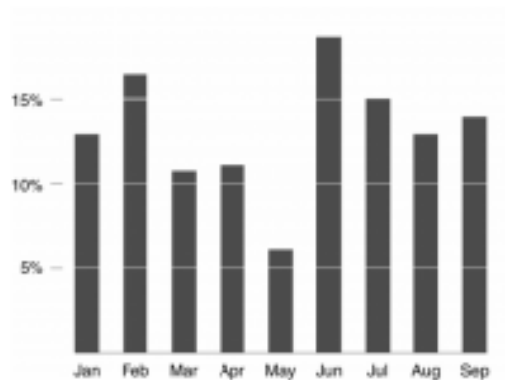
Number of deaths per week from all causes, 2020 vs recent years:  Shading indicates total excess deaths during outbreak



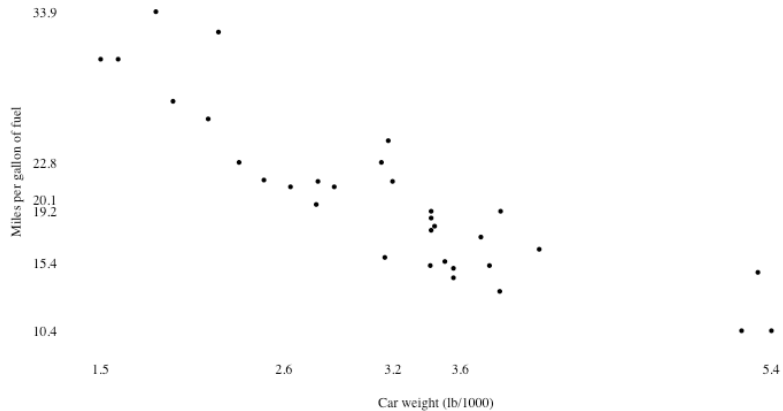
Source: FT analysis of mortality data. Data updated September 25
 FT graphic: John Burn-Murdoch / @burnmurdoch
 © FT

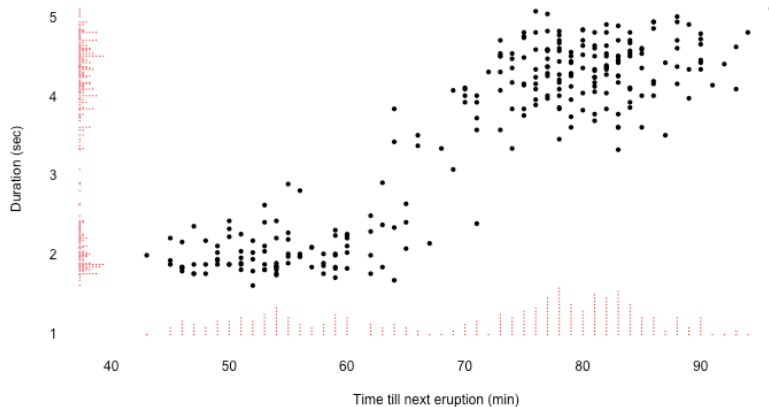


Low Data/Ink

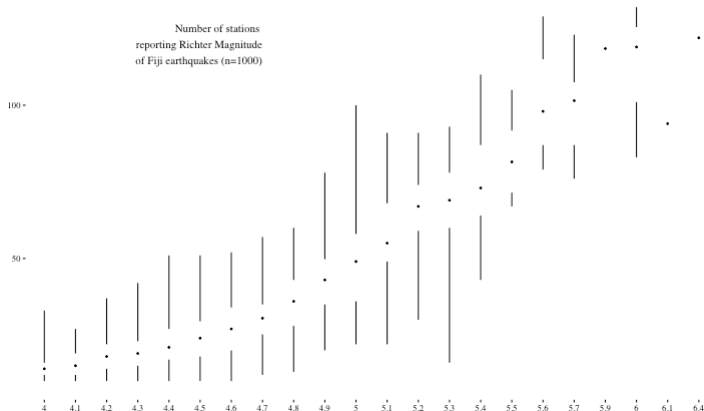


High Data/Ink





```
ggplot(quakes, aes(factor(mag), stations)) +  
  theme_tufte() +  
  geom_tufteboxplot(outlier.colour="transparent") +  
  theme(axis.title=element_blank())
```



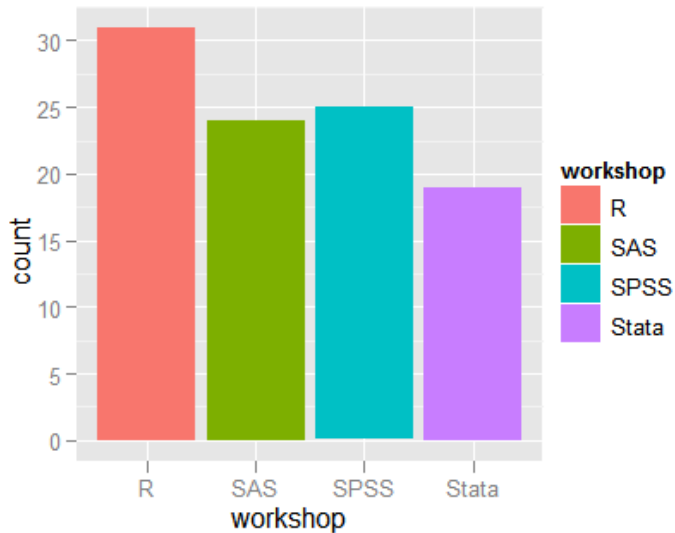
Tufte wisdom

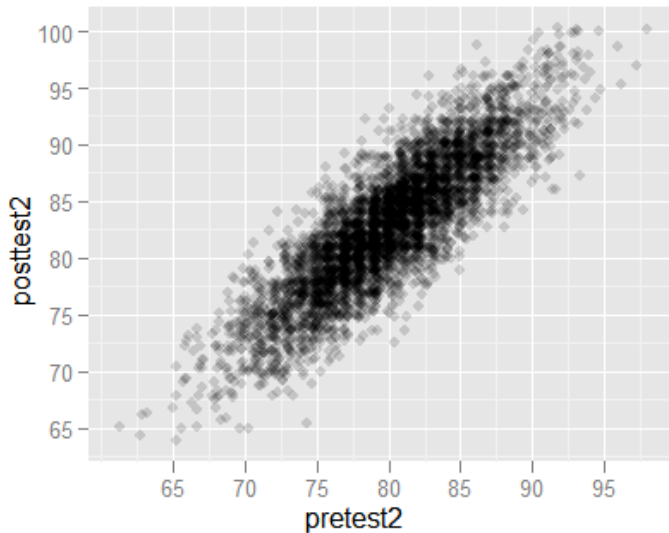
- Tufte's principles are more oriented to communication and can be taken too far
- Better data/ink → display more information without overload;
- Thinking about perception can help you choose better geoms, aesthetics.

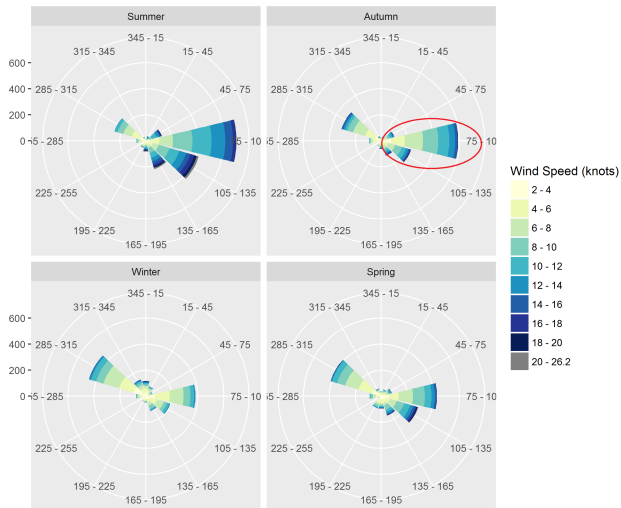
Some practice

Answer these questions:

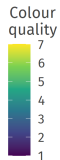
- What are: aesthetics, geom, scale, facets, transformation, coordinate system
- How is data/ink?
- Is perception considered optimally?
- Can you think of questions you can't answer from this plot which are in the data?







and their cut quality



Conclusion

Conclusion

- Data visualization is a huge field;
- Sticking to **basic principles** helps:
 - **Map data** to aesthetics, geoms, scales, facets;
 - Perception research guides choices;
 - **Which comparisons** do I want?
 - Maximize **data-ink** (within reason).