Supervised Learning and Visualization:

Exploratory Data Analysis

Gerko Vink
Erik-Jan van Kesteren
Peter van der Heijden
David Goretzko

Department of Methodology and Statistics

Utrecht Applied Data Science

John Tukey (1915–2000)
Data Scientist patient zero

Inventor of:

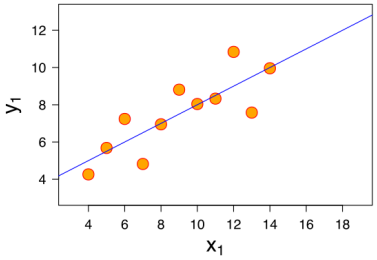
- The boxplot
- The term “exploratory data analysis”
- The Fast Fourier Transform
- “Tukey’s test”
- The word “bit”
- So, so much more (Wikipedia)

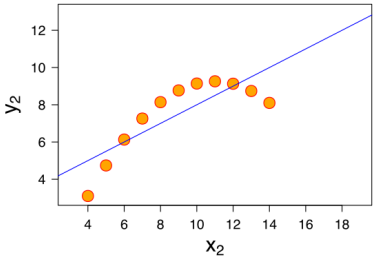
Today: visualization principles, applicable to EDA

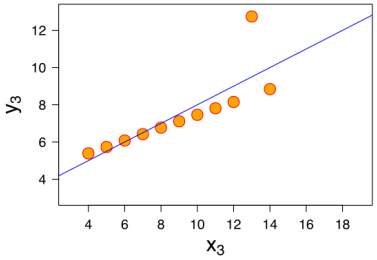
Some data visualization principles

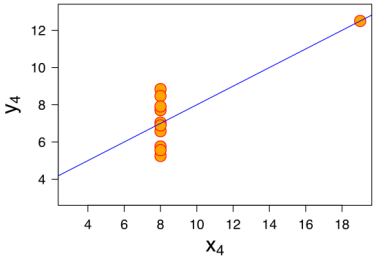
Data visualization

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


y1
x1
4
6
8
10
12
14
16
18


y2
x2
4
6
8
10
12
14
16
18


y3
x3
4
6
8
10
12
14
16
18


y4
x4
4
6
8
10
12
14
16
18

Graphics for data analysis

- The **human retina** can transfer around 10^6 or 10^7 bits per second to the brain;
- **Reading** transfers about 3 words, so $\sim 10^2$ or 10^3 bits/s;
- Potentially (!) visualization is about 4 orders of magnitude more powerful.

How can we leverage the human visual system to analyze data?

Making pictures that help analyze data

- 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

Grammar of graphics (Wickham version)

https://r4ds.had.co.nz/data-visualisation.html

Map raw data to following elements:

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

Additionally, can apply:

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

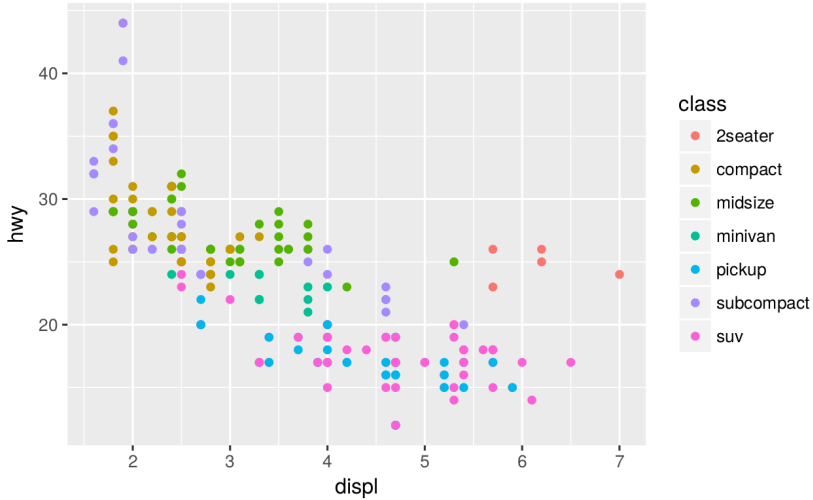
Grammar of graphics (Wickham version)

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

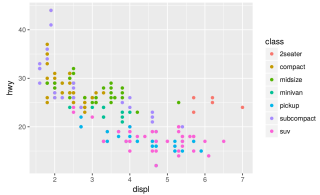
Example data set: cars

```
mpg
#> # 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))
```



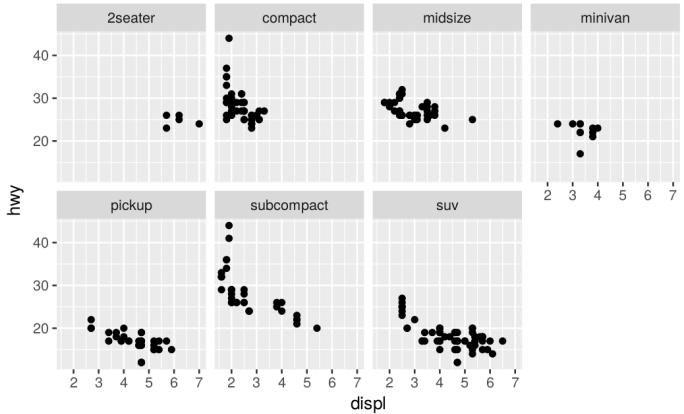
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = class))
```



- 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

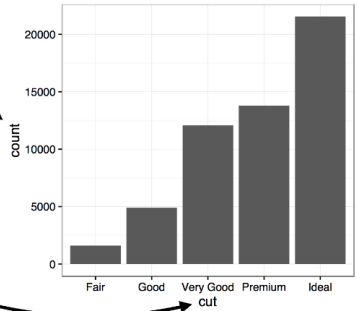
2. **`geom_bar()`** transforms the data with the "count" stat, which returns a data set of cut values and counts.

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.

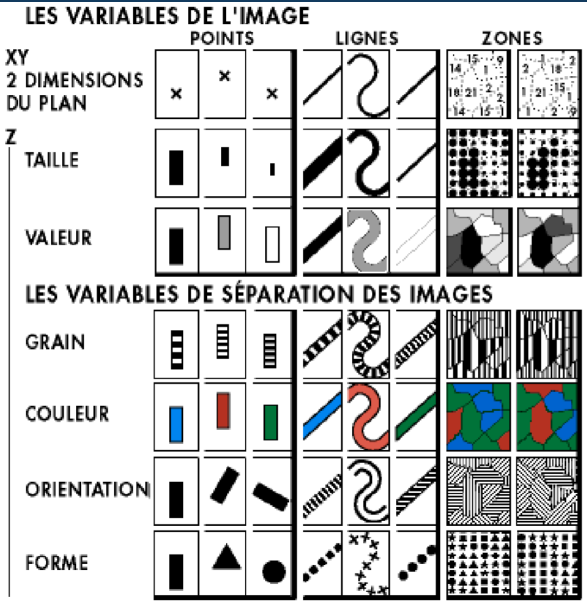
carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
…	…	…	…	…	…	…	…	…	…

stat_count()

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

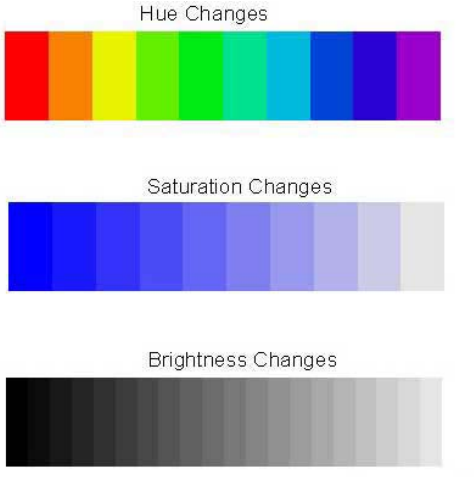


How should I visualize my data/ analysis results?



Jacques Bertin (1967) Sémiologie graphique

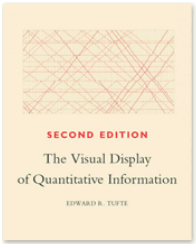
Color: hue-saturation-brightness (HSB)

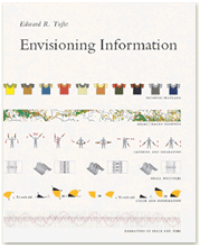


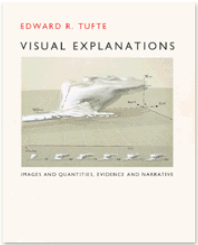
Mackinlay's ranking of encodings

Quantitative	**Ordinal**	**Nominal**
Position	Position	Position
Length	Density	Color hue
Angle	Color saturation	Texture
Slope	Color hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Color saturation
Color saturation	Length	Shape
Color hue	Angle	Length
Texture	Slope	Angle
Connection	Area	Slope
Containment	Volume	Area
Shape	Shape	Volume

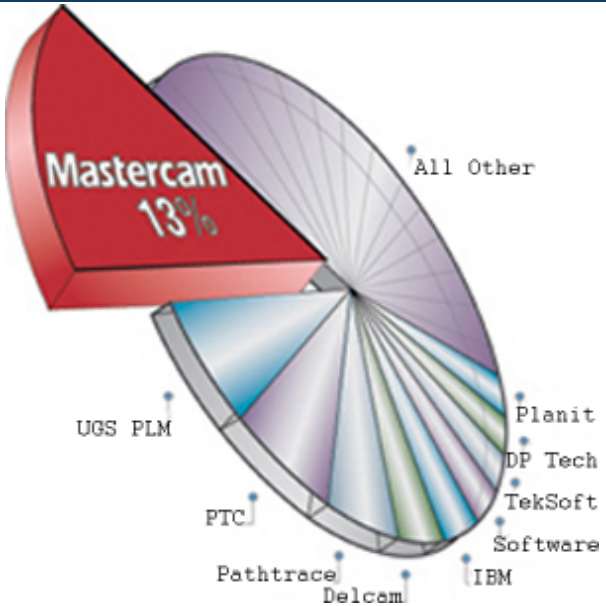
Some (distilled) principles from Tufte



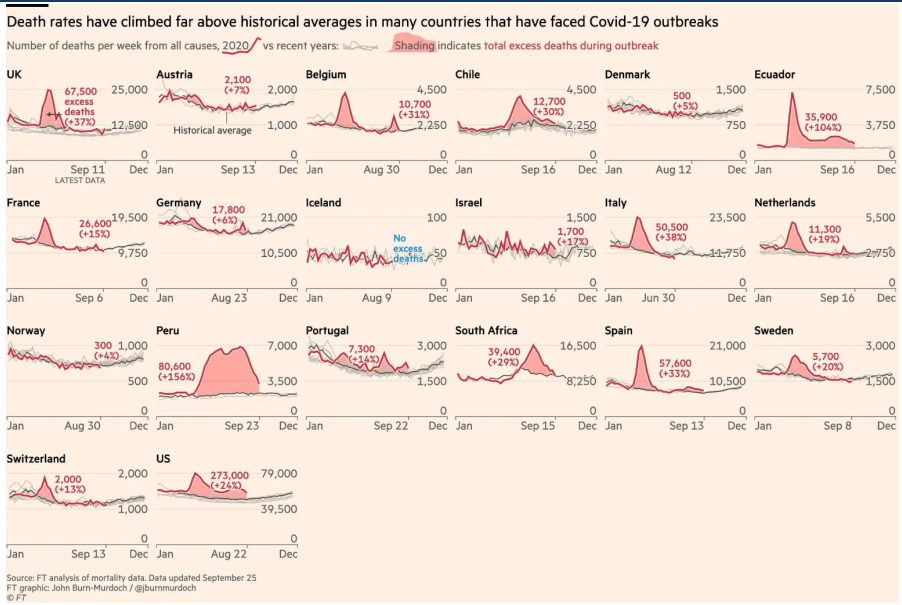


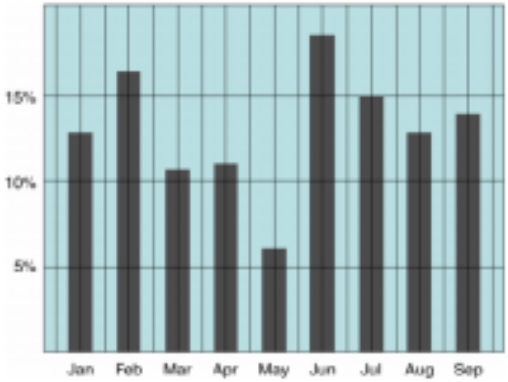


- Ask how data maps to perception
- Ask which comparisons you want, guide eye to those
- Maximize **data-to-ink ratio**
- Present more data (but without losing interpretability)
- (Remember narrative)

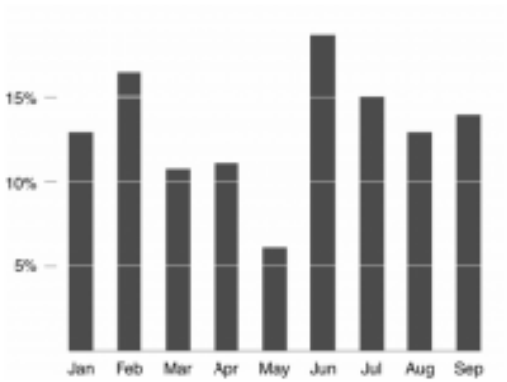

Mastercam
13%
All Other
UGS PLM
Planit
DP Tech
TekSoft
Software
PTC
Pathtrace
Delcam
IBM


DIAMONDS *WERE* A GIRL'S BEST FRIEND
Average price of a one-carat D-flawless
$60,000
$50,000
$40,000
$30,000
$20,000
1978
1979
1980
1981
1982
DIAMONDS WERE A GIRL'S BEST FRIEND
Average price of a one-carat D-flawless
$60,000
$50,000
$40,000
$30,000
$20,000
1978
1979
1980
1981
1982

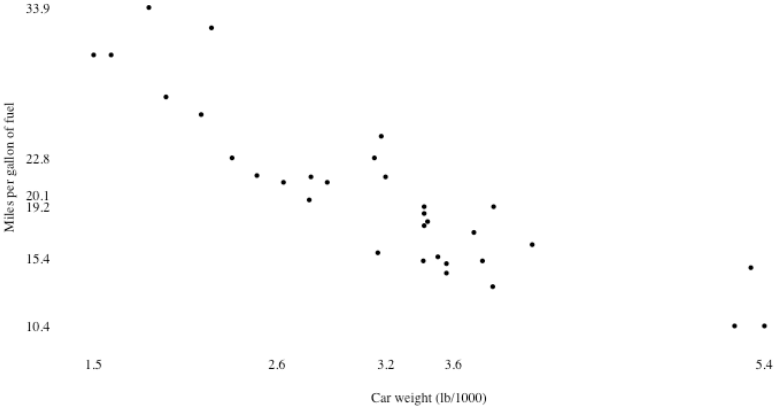

Death rates have climbed far above historical averages in many countries that have faced Covid-19 outbreaks
Number of deaths per week from all causes, 2020 vs recent years:
Shading indicates total excess deaths during outbreak
UK
67,500 excess deaths (+37%)
25,000
12,500
0
Jan
Sep 11
Dec
LATEST DATA
Austria
2,100 (+7%)
Historical average
2,000
1,000
0
Jan
Sep 13
Dec
Belgium
10,700 (+31%)
4,500
2,250
0
Jan
Aug 30
Dec
Chile
12,700 (+30%)
4,500
2,250
0
Jan
Sep 16
Dec
Denmark
500 (+5%)
1,500
750
0
Jan
Aug 12
Dec
Ecuador
35,900 (+104%)
7,500
3,750
0
Jan
Sep 16
Dec
France
26,600 (+15%)
19,500
9,750
0
Jan
Sep 6
Dec
Germany
17,800 (+6%)
21,000
10,500
0
Jan
Aug 23
Dec
Iceland
No excess deaths
100
50
0
Jan
Aug 9
Dec
Israel
1,700 (+17%)
1,500
750
0
Jan
Sep 16
Dec
Italy
50,500 (+38%)
23,500
11,750
0
Jan
Jun 30
Dec
Netherlands
11,300 (+19%)
5,500
2,750
0
Jan
Sep 16
Dec
Norway
300 (+4%)
1,000
500
0
Jan
Aug 30
Dec
Peru
80,600 (+156%)
7,000
3,500
0
Jan
Sep 23
Dec
Portugal
7,300 (+14%)
3,000
1,500
0
Jan
Sep 22
Dec
South Africa
39,400 (+29%)
16,500
8,250
0
Jan
Sep 15
Dec
Spain
57,600 (+33%)
21,000
10,500
0
Jan
Sep 13
Dec
Sweden
5,700 (+20%)
3,000
1,500
0
Jan
Sep 8
Dec
Switzerland
2,000 (+13%)
2,000
1,000
0
Jan
Sep 13
Dec
US
273,000 (+24%)
79,000
39,500
0
Jan
Aug 22
Dec
Source: FT analysis of mortality data. Data updated September 25
FT graphic: John Burn-Murdoch / @jburnmurdoch
© FT

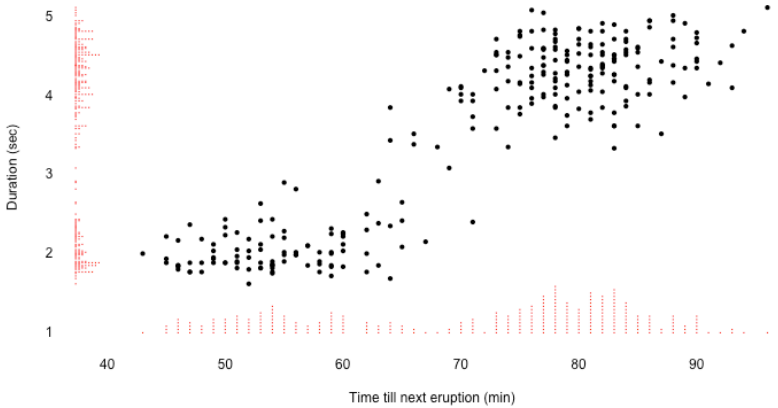


Low Data/Ink



High Data/Ink


33.9
22.8
20.1
19.2
15.4
10.4
Miles per gallon of fuel
1.5
2.6
3.2
3.6
5.4
Car weight (lb/1000)


Duration (sec)
5
4
3
2
1
40
50
60
70
80
90
Time till next eruption (min)

Tufte wisdom

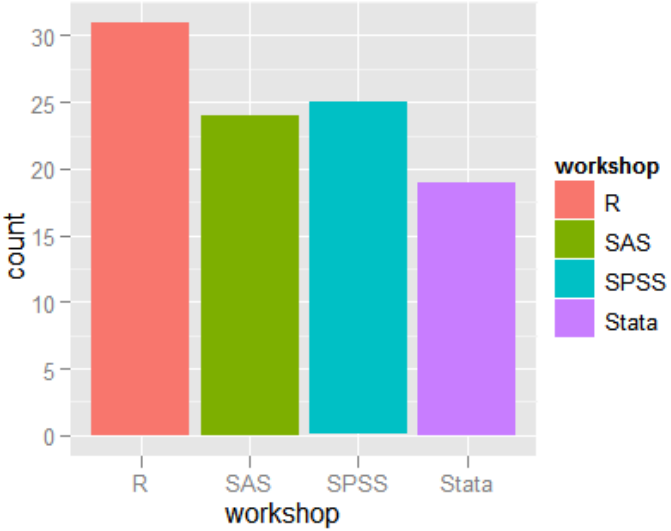
- Tufte's principles are more oriented to communication and can be taken too far

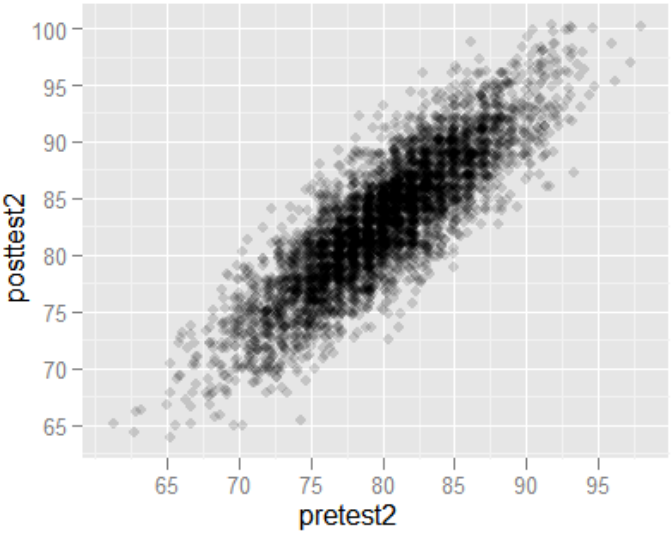
- Better data/ink $\rightarrow$ display more information without overload;
- Thinking about perception can help you choose better geoms, aesthetics.

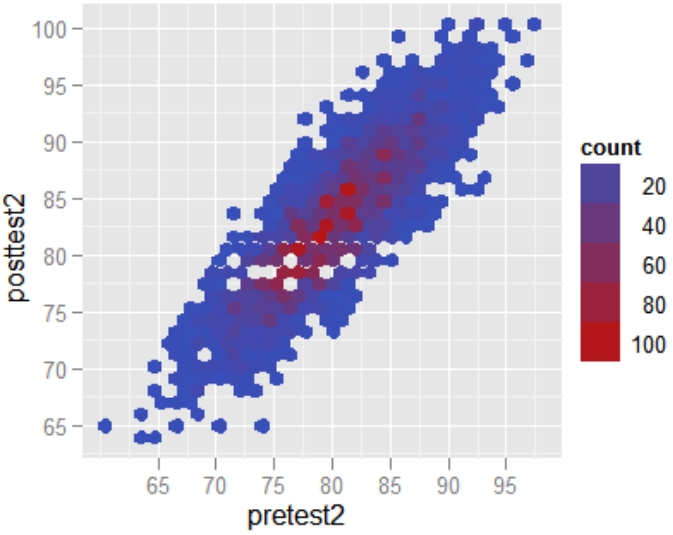
Some practice

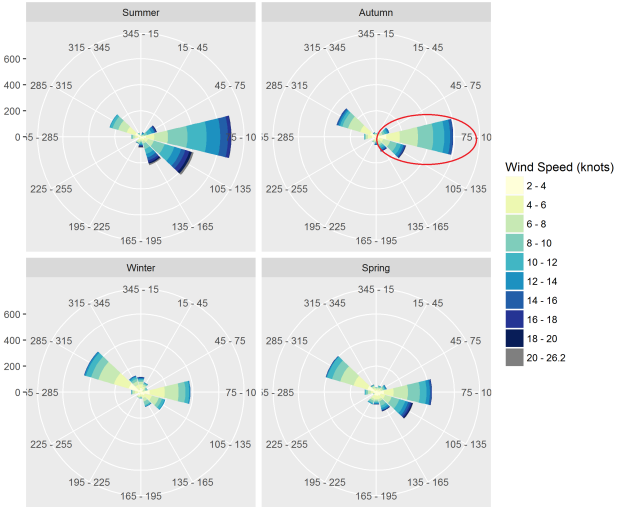
Answer these questions:

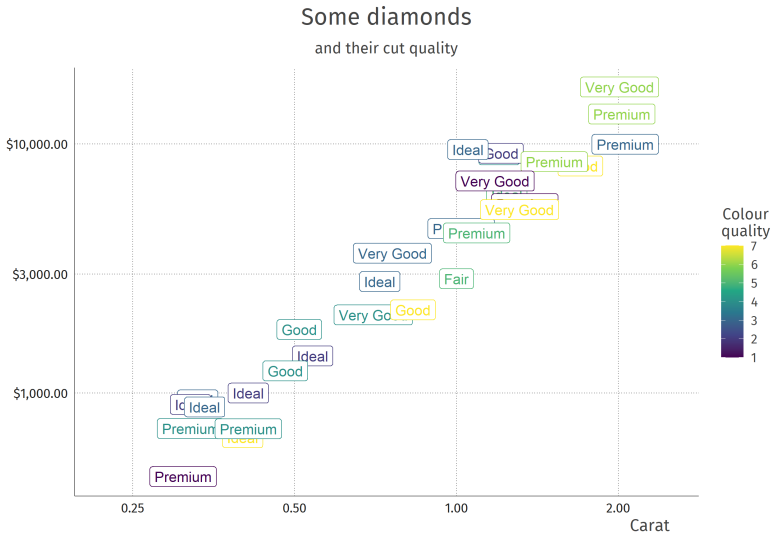
- What are: aesthetics, geom, scale, facets, transformation, coordinate system
- How is data/ink?
- Is perception considered optimally?
- Can you think of questions you can't answer from this plot which are in the data?


30
25
20
15
10
5
0
count
R
SAS
SPSS
Stata
workshop
workshop
R
SAS
SPSS
Stata


posttest2
100
95
90
85
80
75
70
65
65
70
75
80
85
90
95
pretest2


posttest2
100
95
90
85
80
75
70
65
65
70
75
80
85
90
95
pretest2
count
20
40
60
80
100


Summer
Autumn
Winter
Spring
345 - 15
15 - 45
45 - 75
75 - 105
105 - 135
135 - 165
165 - 195
195 - 225
225 - 255
285 - 315
315 - 345
600
400
200
0
Wind Speed (knots)
2 - 4
4 - 6
6 - 8
8 - 10
10 - 12
12 - 14
14 - 16
16 - 18
18 - 20
20 - 26.2


Some diamonds
and their cut quality
$10,000.00
$3,000.00
$1,000.00
0.25
0.50
1.00
2.00
Carat
Colour quality
7
6
5
4
3
2
1
Very Good
Premium
Premium
Ideal
Premium
Very Good
Very Good
Premium
Very Good
Ideal
Fair
Good
Good
Ideal
Good
Ideal
Premium
Premium
Premium

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).
- There is no ’one solution fits all’ approach!