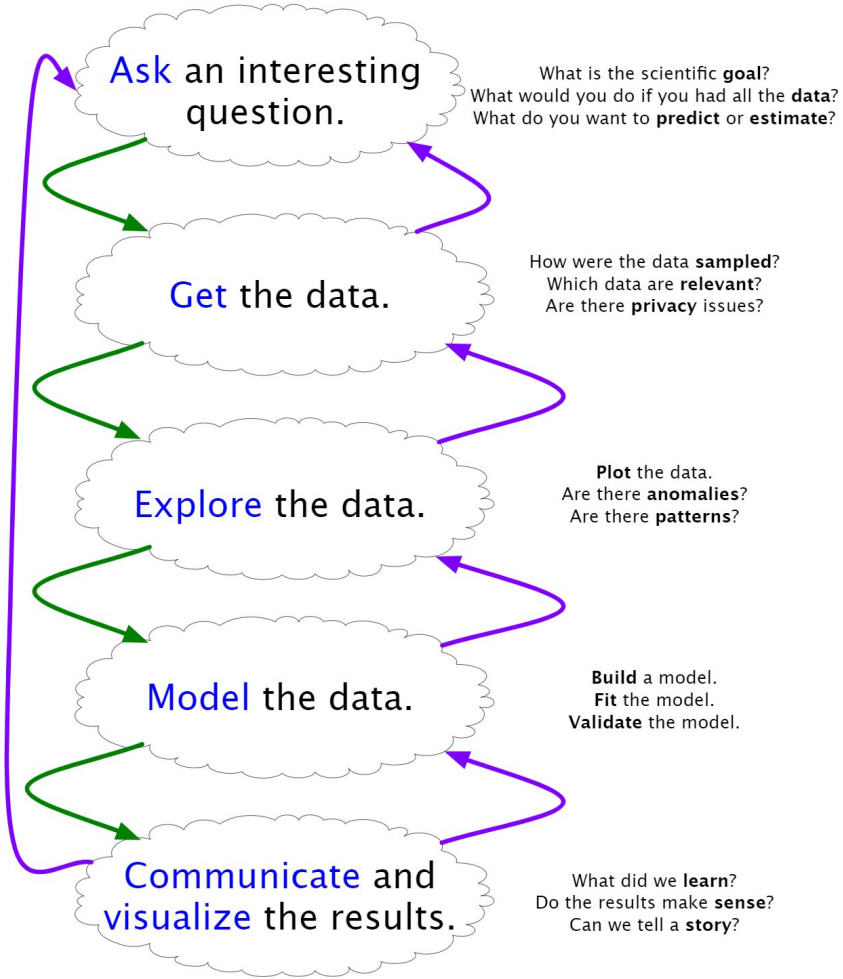
Practical Data Science
Week 2 - Lecture 2

Exploratory Data Analysis & Effective Visualizations


Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?

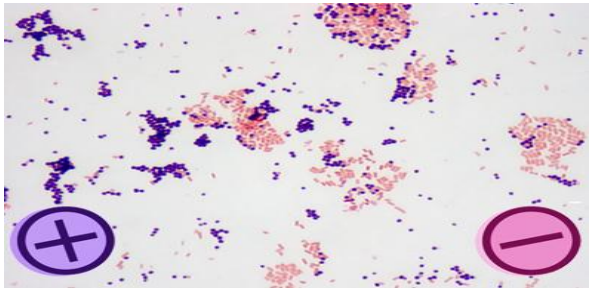
Data exploration

Not always sure what we are looking for (until we find it)

Example: Antibiotics

Will Burtin

Data



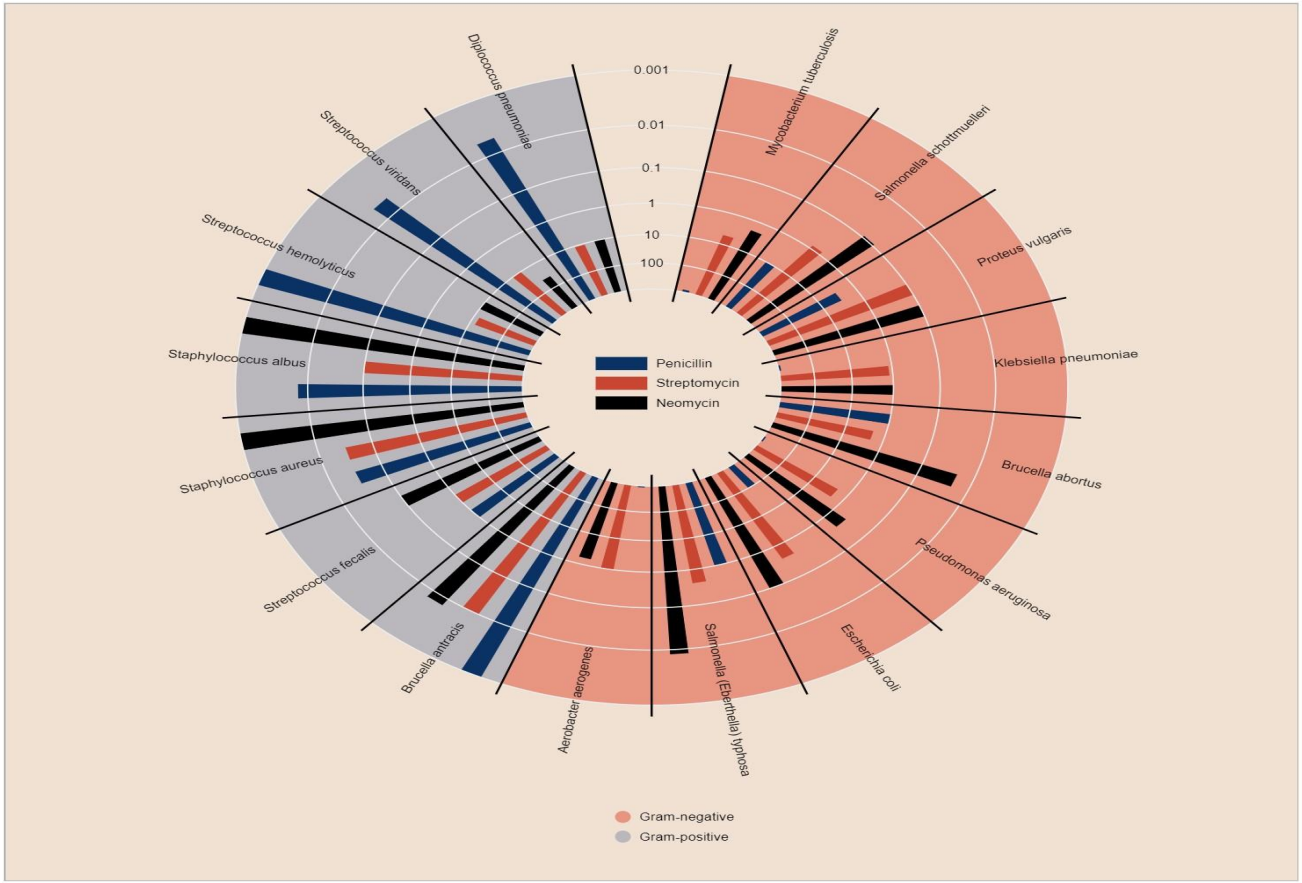
Genus, Species

Table 1: Burtin's data.

Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	850	1.2	1	negative
Mycobacterium *tuberculosis*	800	5	2	negative
Proteus *vulgaris*	3	0.1	0.1	negative
Pseudomonas *aeruginosa*	850	2	0.4	negative
Salmonella (Eberthella) *typhosa*	1	0.4	0.008	negative
Salmonella *schottmuelleri*	10	0.8	0.09	negative
Staphylococcus *albus*	0.007	0.1	0.001	positive
Staphylococcus *aureus*	0.03	0.03	0.001	positive
Streptococcus *fecalis*	1	1	0.1	positive
Streptococcus *hemolyticus*	0.001	14	10	positive
Streptococcus *viridans*	0.005	10	40	positive

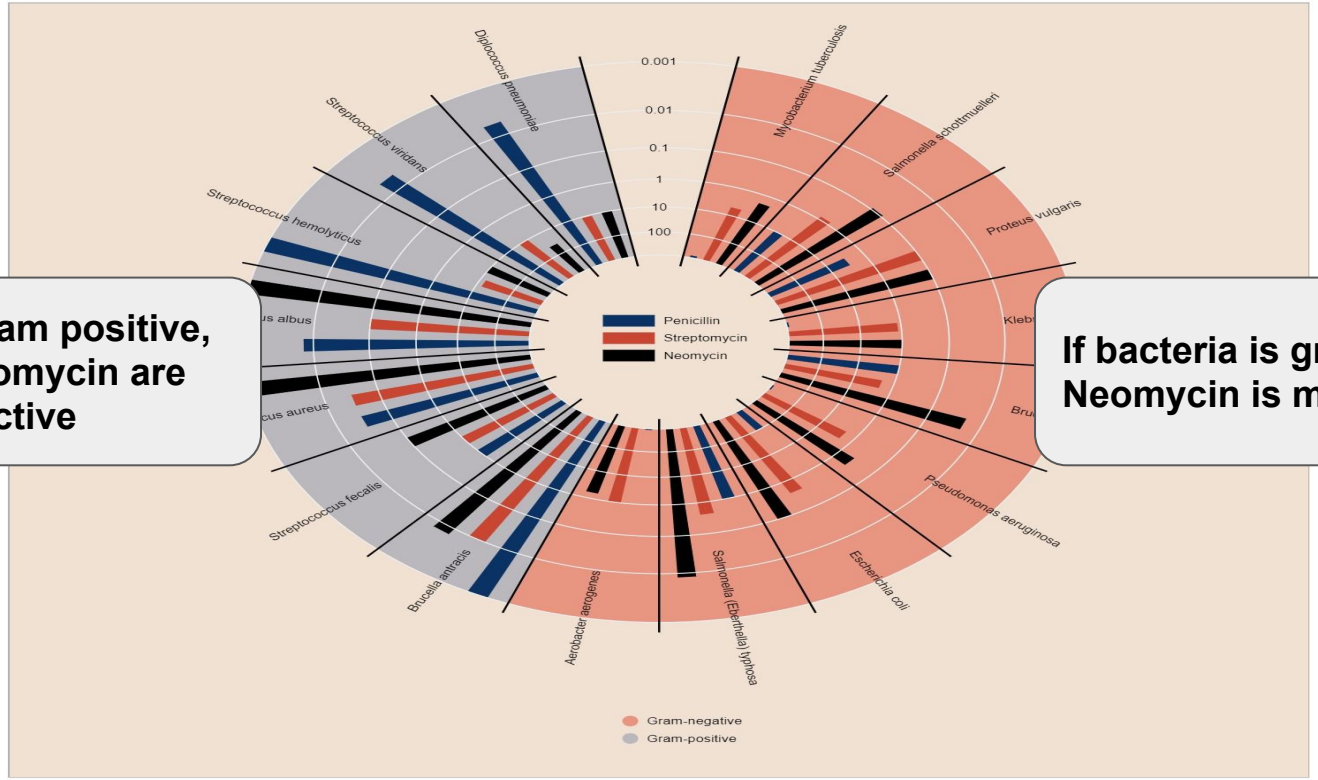
Burtin's Antibiotics

Gram Positive



Gram Negative

Burtin's Antibiotics



If bacteria is gram positive, Penicillin & Neomycin are most effective

If bacteria is gram negative, Neomycin is most effective.

Exploratory Data Analysis

“The greatest value of a picture is when it forces us to notice what we never expected to see.”

John Tukey

Visualization

To convey information through graphical representations of data

Visualization Goals

Communicate (Explanatory)

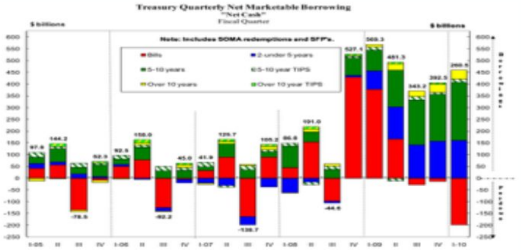
Present data and ideas
Explain and inform
Provide evidence and support
Influence and Persuade

Analyze (Exploratory)

Explore the data
Assess a situation
Determine how to proceed
Decide what to do

Effective visualizations

Not Effective...



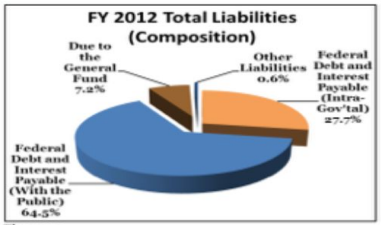
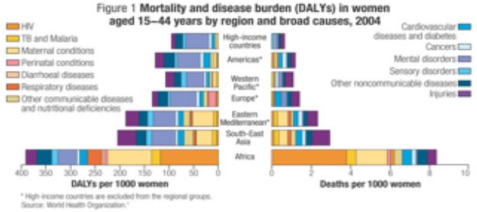
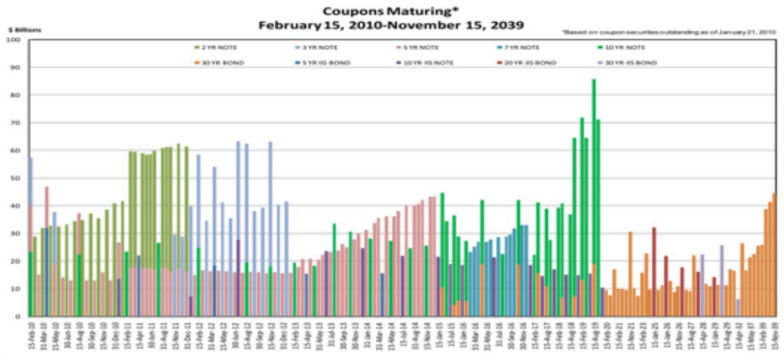


Figure 10





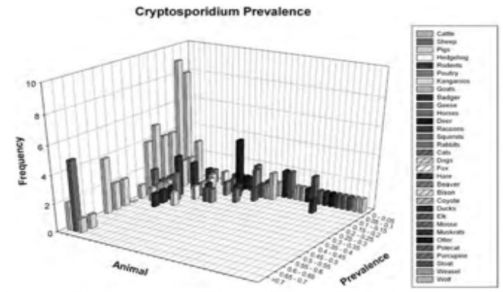


Figure 5.2 Mean prevalence rates of *Cryptosporidium* oocysts by animal species.

Sources: US Treasury and WHO reports

Effective Visualizations

1. Have graphical integrity
2. Keep it simple
3. Use the right display
4. Use color strategically

Graphical Integrity

Bonus Assignment

Chart #1:

Issues:

- Tell a Story
  - The vertical axis isn't labeled. We don't know the unit.
- Graphical Integrity
  - The vertical axis does not start from zero
- Graphical Complexity
  - Horizontal and vertical lines are unnecessary (Chartjunk)

Bonus Assignment

Chart #1:

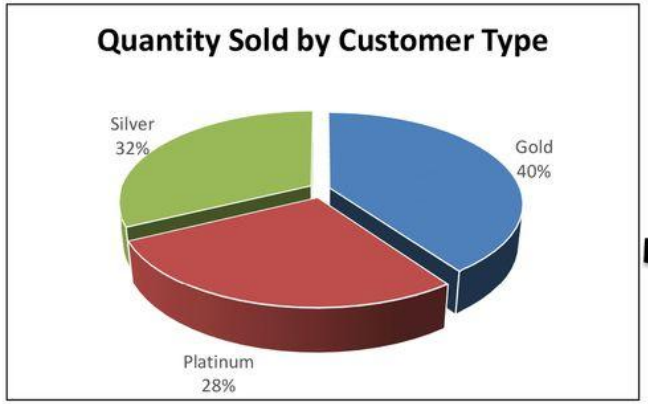
Issues:

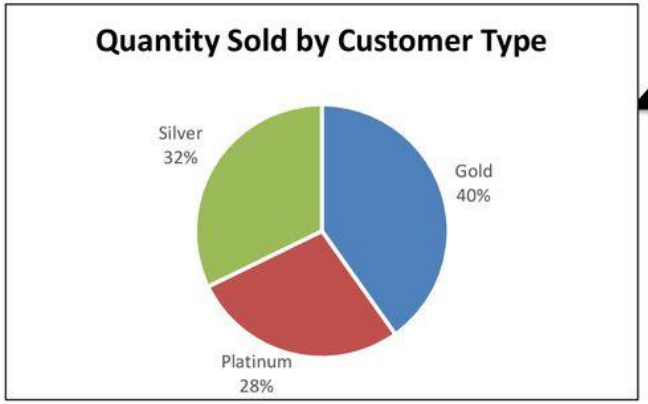
- Tell a Story
  - The vertical axis isn't labeled. We don't know the unit.
- Graphical Integrity
  - The vertical axis does not start from zero
- Graphical Complexity
  - Horizontal and vertical lines are unnecessary (Chartjunk)

This also works

Bonus Assignment

Chart #2:



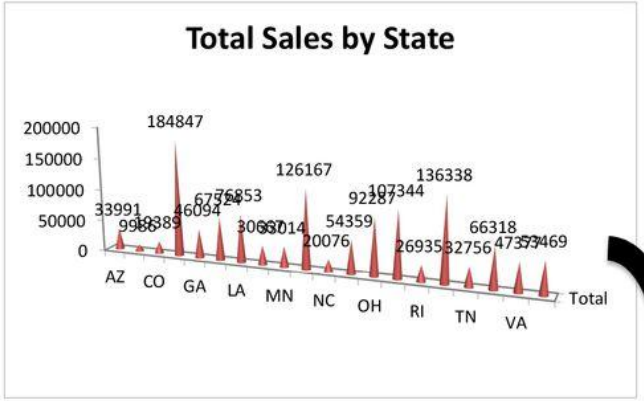


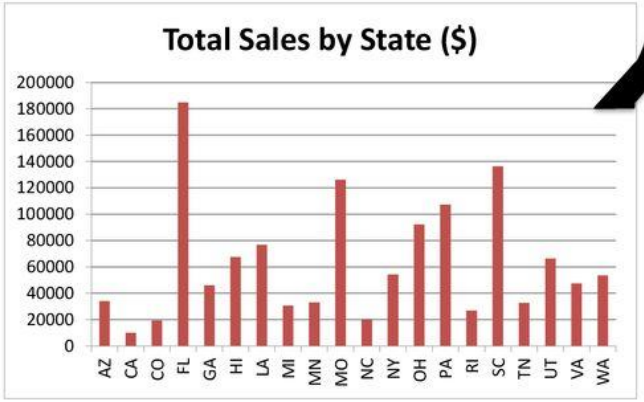
Issues:

- Graphical Integrity
  - The 3D chart makes it difficult to compare the sizes
- Graphical Complexity
  - The 3D chart requires more ink (Chartjunk)

Bonus Assignment

Chart #3:

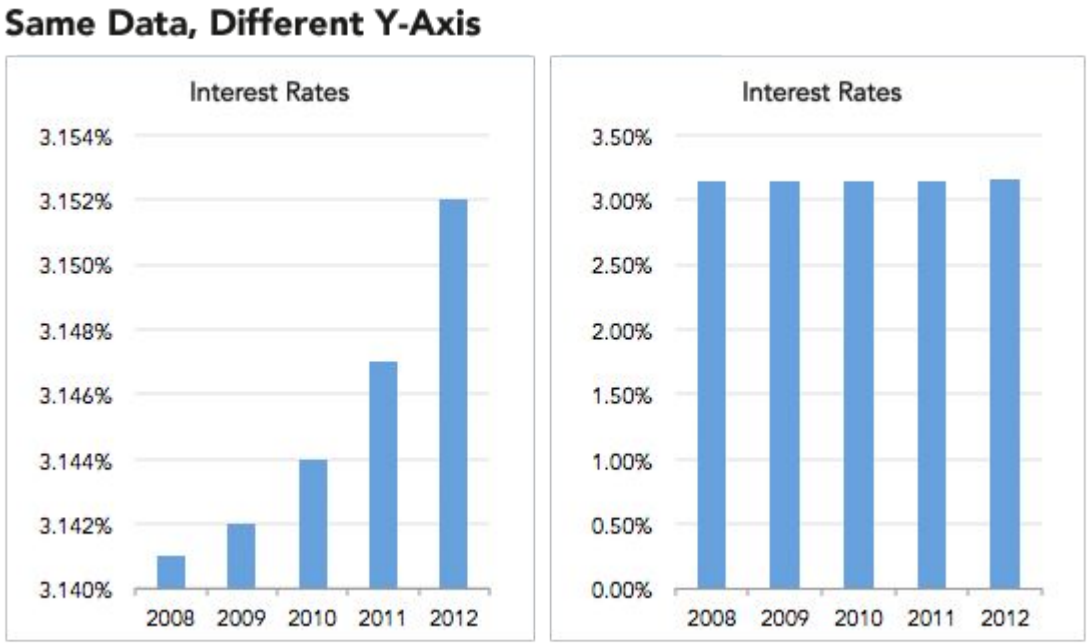




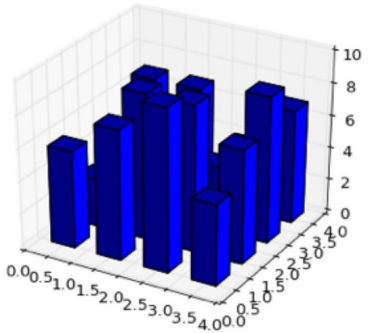
Issues:

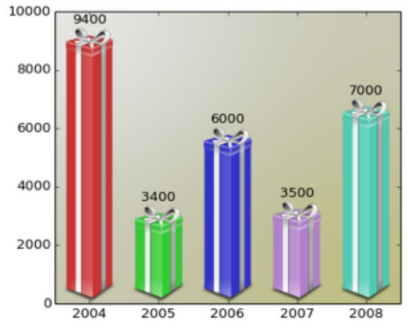
- Tell a Story
  - Vertical axis isn't labeled. We don't know the units
  - Because there are many states to compare, horizontal lines may be helpful
- Graphical Integrity
  - The 3D chart makes it difficult to compare sizes
  - The cone-shaped bars make it even harder to compare sizes
- Graphical Complexity
  - The 3D chart requires more ink (Chartjunk)
  - The number labels are unnecessary

Scale distortion

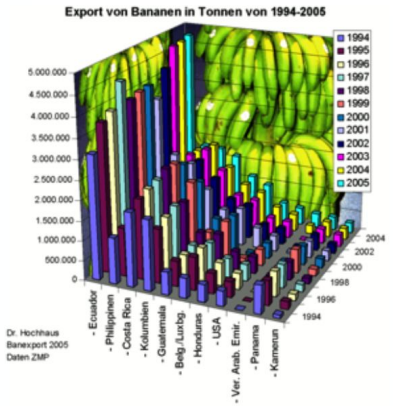


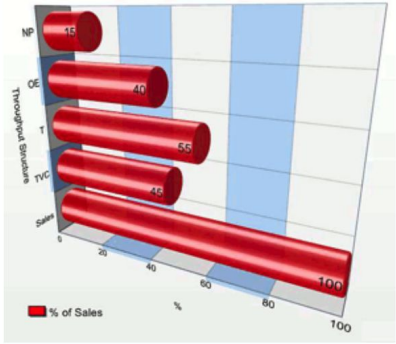
Don’t!





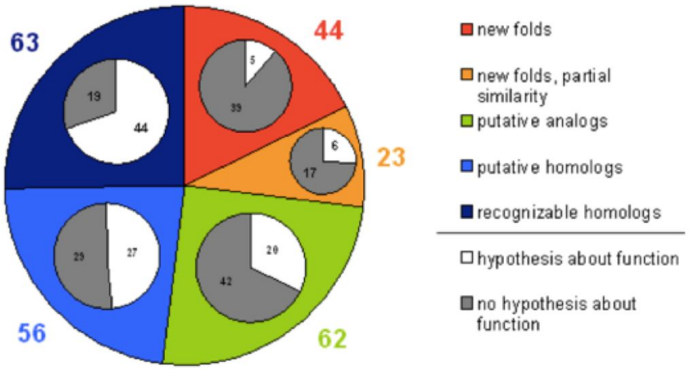
matplotlib gallery

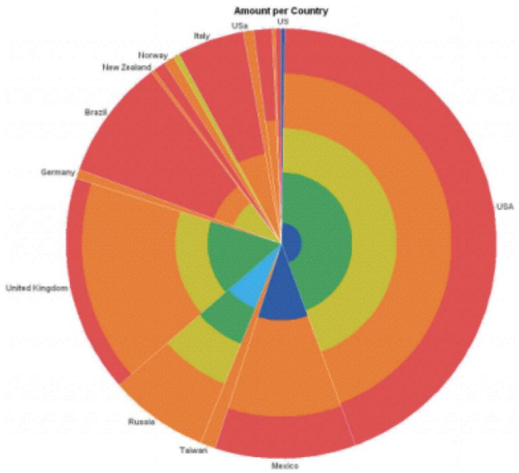




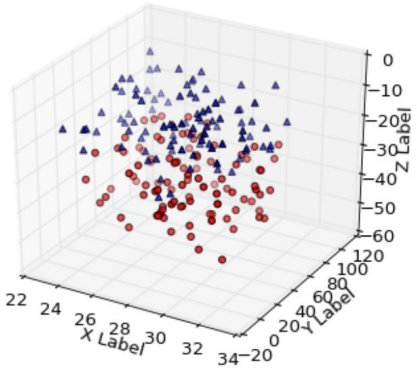
Excel Charts Blog

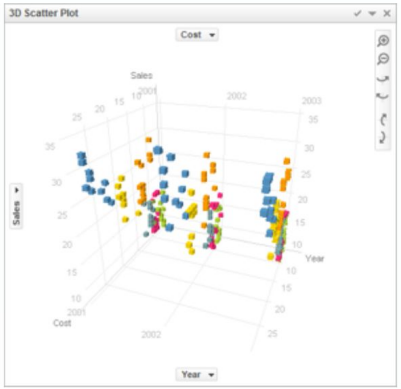
Don’t!

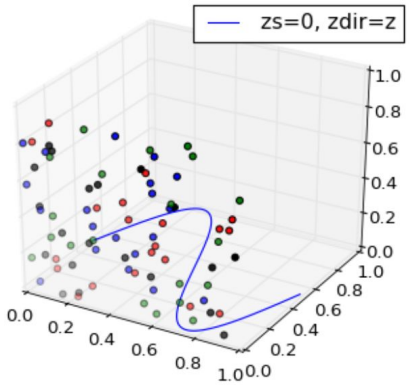

63
44
23
56
62
5
39
6
17
19
44
29
27
20
42
new folds
new folds, partial similarity
putative analogs
putative homologs
recognizable homologs
hypothesis about function
no hypothesis about function

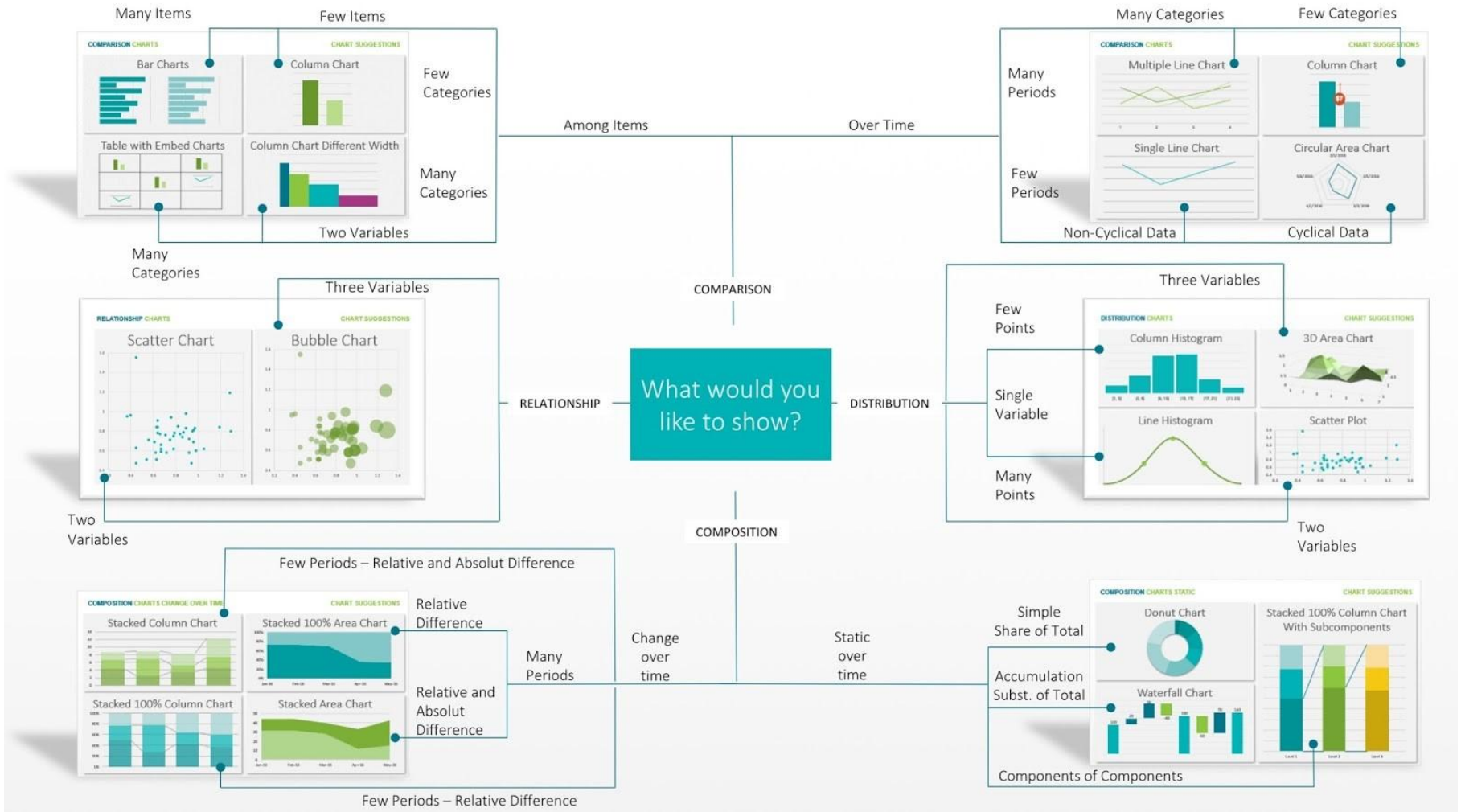

Amount per Country
US
USa
Italy
Norway
New Zealand
Brazil
Germany
United Kingdom
USA
Russia
Taiwan
Mexico

Don’t!


X Label
Y Label
Z Label


3D Scatter Plot
Cost
Sales
Year


zs=0, zdir=z


What would you like to show?
COMPARISON
Among Items
Over Time
Many Items
Few Items
Few Categories
Many Categories
Two Variables
Many Categories
Many Periods
Few Periods
Non-Cyclical Data
Cyclical Data
RELATIONSHIP
Three Variables
Two Variables
DISTRIBUTION
Three Variables
Few Points
Single Variable
Many Points
Two Variables
COMPOSITION
Change over time
Few Periods – Relative and Absolut Difference
Relative Difference
Many Periods
Relative and Absolut Difference
Few Periods – Relative Difference
Static over time
Simple Share of Total
Accumulation Subst. of Total
Components of Components

Exercise

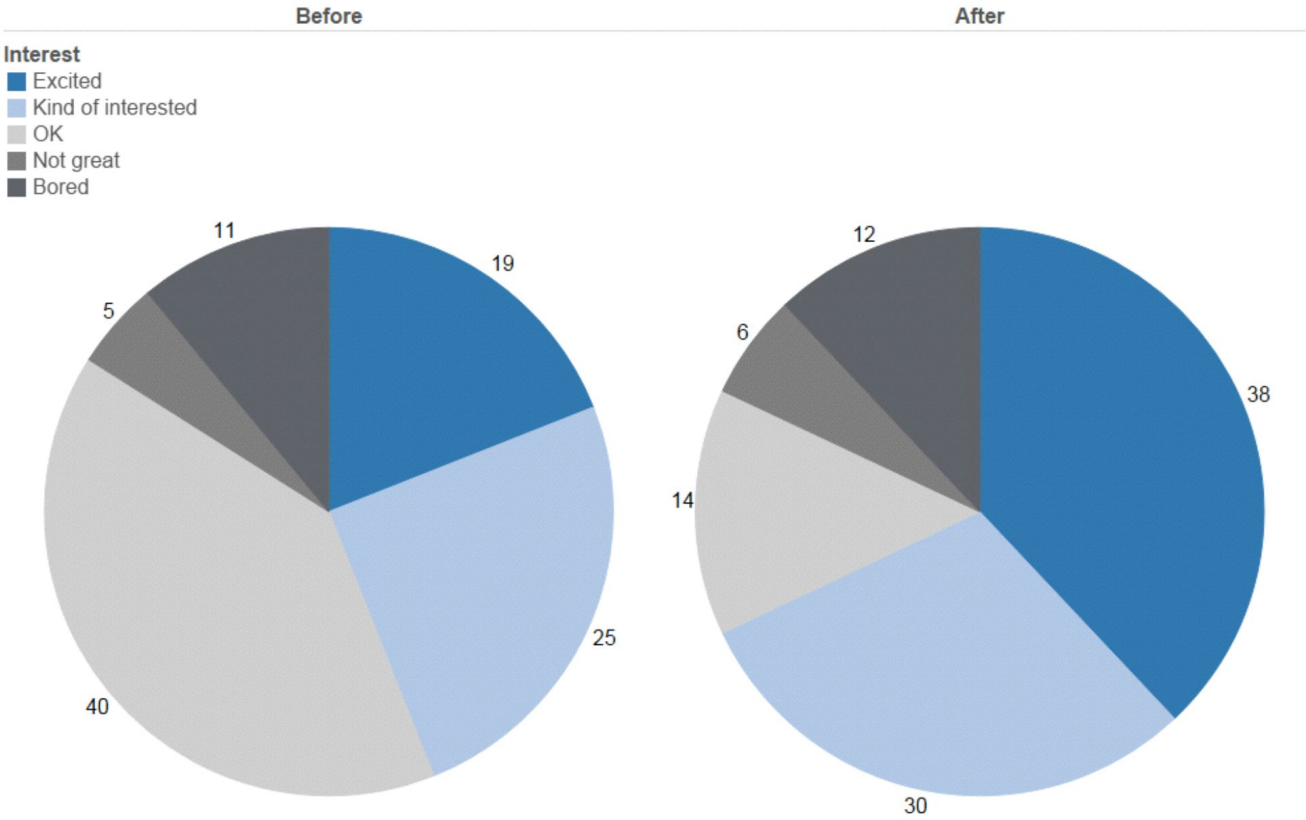
How do you feel about doing science?

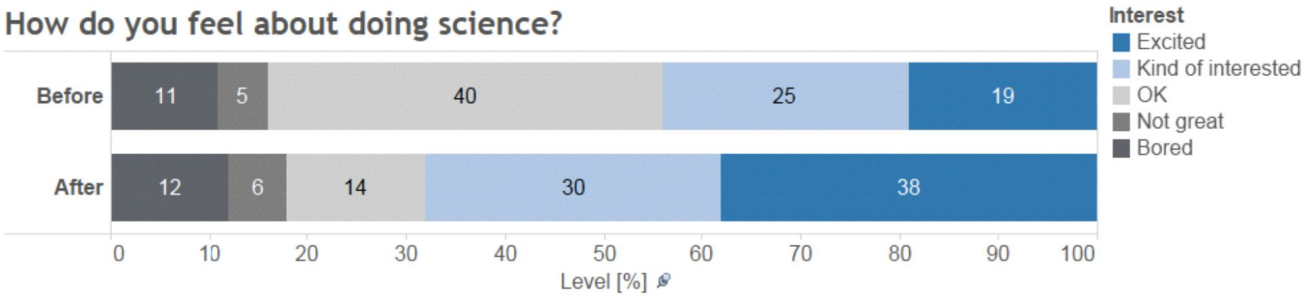
Table

Interest	Before	After
Excited	19	38
Kind of interested	25	30
OK	40	14
Not great	5	6
Bored	11	12

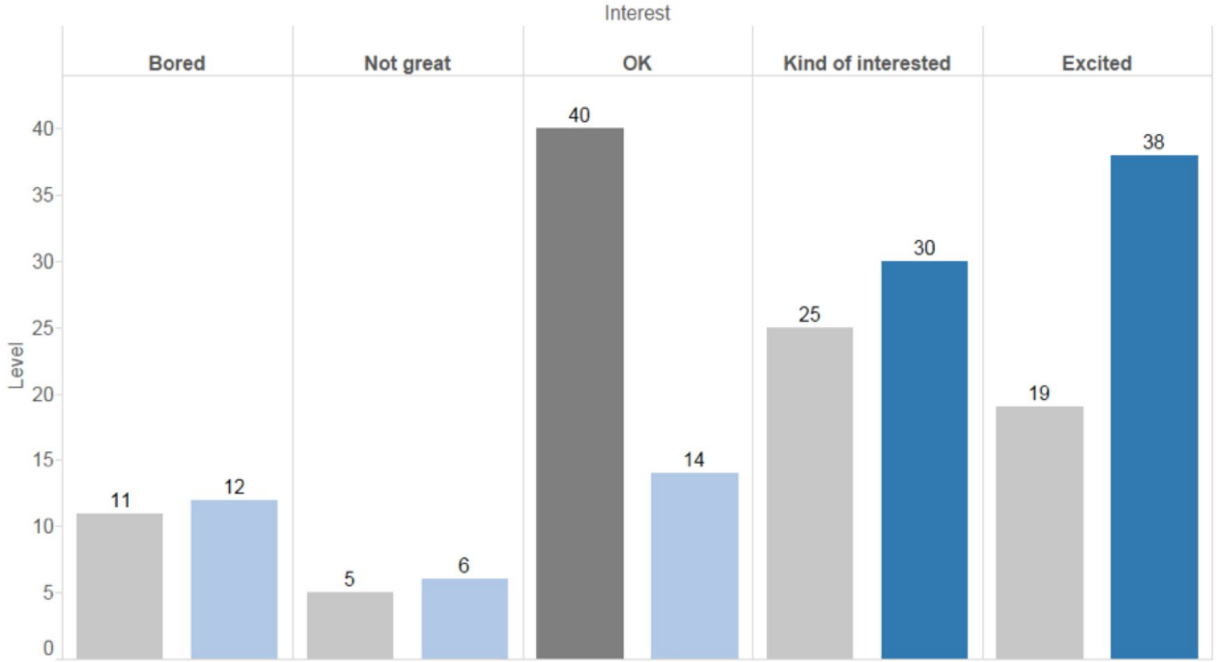
Data courtesy of Cole Nussbaumer

How do you feel about doing science?



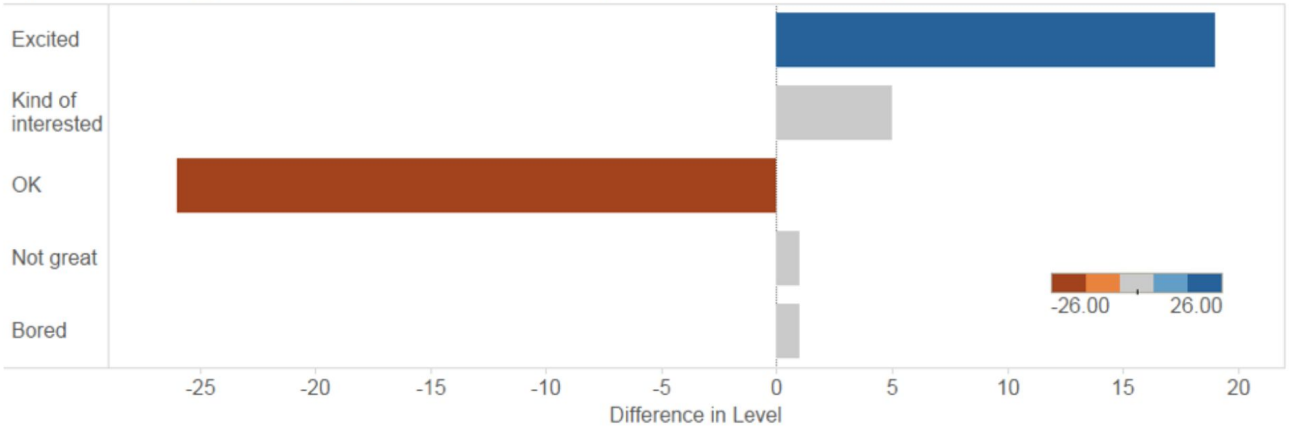

How do you feel about doing science?
Before
11
5
40
25
19
After
12
6
14
30
38
0
10
20
30
40
50
60
70
80
90
100
Level [%]
Interest
Excited
Kind of interested
OK
Not great
Bored

How do you feel about doing science?

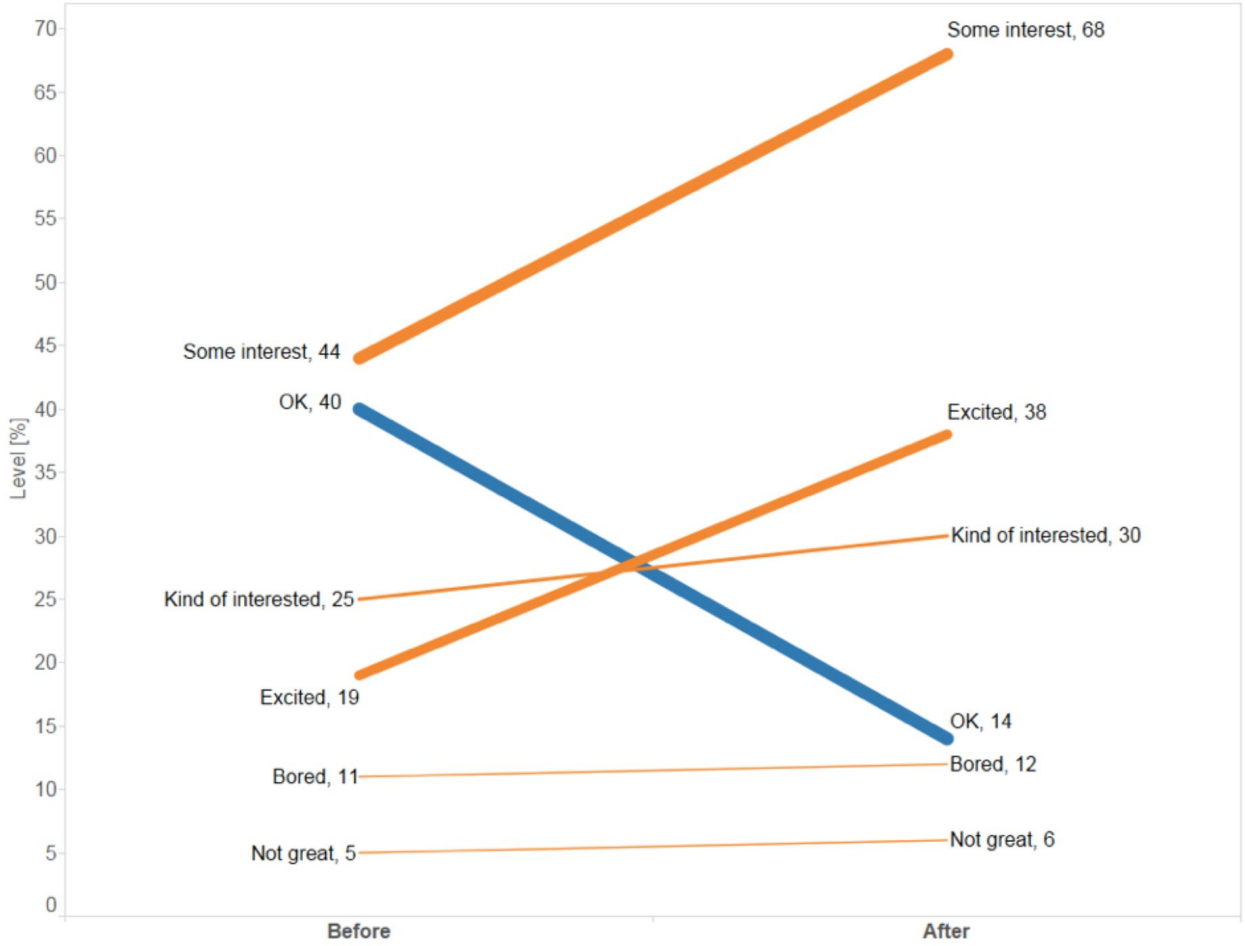


Before the program, the majority of children felt just ***OK*** about science. After the program, more children were ***Kind of interested*** and ***Excited*** about science.

Opinion change to the question: How do you feel about doing science?



How do you feel about doing science?



After the pilot program,

68%

of kids expressed interest towards science,
compared to 44% going into the program.