S-109a: Data Science

Effective Exploratory Data Analysis and Visualization

Pavlos Protopapas, Kevin Rader, Rahul Dave

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**?

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.
VISUALIZE 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**?

Anscombe's Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.

Unstructured Quartet

Each dataset here also has the same summary statistics. However, they are not *clearly different* or *visually distinct*.

X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

https://www.autodeskresearch.com/publications/samestats

Example: Antibiotics

Will Burtin, 1951

Data

Genus, *Species*

Table 1: Burtin's data.

Min. Inhibitory Concentration [ml/g]

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

What Questions?

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

How effective are the drugs?

Gram Positive

Gram Negative

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

If bacteria is gram negative, Neomycin is most effective

M. Bostock, Protovis after W. Burtin, 1951

How do the bacteria compare?

Not a streptococcus!
(realized ~30 years later)

Really a streptococcus!
(realized ~20 years later)

Wainer & Lysen, "That's funny..."
American Scientist, 2009
Adapted from Brian Schmotzer

How do the bacteria compare?

Wainer & Lysen, "That's funny..."
American Scientist, 2009

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 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

Communicate

Explore

75
70
65
60
55
100
150
200
0.2
0.3
0.4
0.5
0.6
0.8
0.9

1000
800
600
400
200
0
Cost
Slope
Intercept

Action
Adult
Adventure
Animation
Diography
Comedy
Crime
Drama
Family
Fantasy
Film-Noir
History
Horror
Music
Musical
Mystery
News
Reality-TV
Romance
Sci Fi
Sport
Thriller
War
Western
Score

scatter plot between log(tau) and theta[0]
log(tau)
theta[0]
4
2
0
−2
−4
−6
−20
−10
0
10
20

1
0
-1
-2
log(τ1)
β11
β12
β13
β14
β15
β16

EDA Workflow

1. **Build** a DataFrame from the data (ideally, put all data in this object)
2. **Clean** the DataFrame. It should have the following properties
   - Each row describes a single object
   - Each column describes a property of that object
   - Columns are numeric whenever appropriate
   - Columns contain atomic properties that cannot be further decomposed
3. Explore **global properties**. Use histograms, scatter plots, and aggregation functions to summarize the data.
4. Explore **group properties**. Use groupby and small multiples to compare subsets of the data.

Viz options

- Pandas Visualization module
- Matplotlib
- Seaborn
- Above 3 are inter-mixable
- Be lazy (to an extent…)
- Other options: Bokeh, Vega, Vincent, Altair

Cars Dataset

	name	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	maker
0	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4	Mazda
1	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4	Mazda
2	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1	Datsun
3	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1	Hornet
4	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2	Hornet

Basic Pandas/matplotlib

Can set limits, tick styles, scales, add lines, annotations, titles, legends

Seaborn provides a different visual style and lots of canned plots.

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 EDA Viz

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

1. Graphical Integrity

Yale Daily News | THE OLDEST COLLEGE DAILY ESTABLISHED 1878

Yale Summer Session
Over 200 full-credit courses.
Same Veritas. More Lux.
June 4 – July 6 , July 9 – Aug 10
2012 experience Yale

CHART YALE GRADUATES' MAJORS, CLASS OF 2011

September 13, 2011

Facebook Recommendations

Shake Shack to open in New Haven
277 people recommend this.

Popular anti-religion creates false dichotomy
15 people recommend this.

Friends remember Foucher LAW '14
10 people recommend this.

AIDS activist speaks about documentary film
8 people recommend this.

Panel outlines changes in hip-hop
30 people recommend this.

Facebook social plugin

Advertisement

Featured Jobs

Graphical Integrity

Flowing Data

Scale Distortions

JOB LOSS BY QUARTER
15 MIL
13.5 MIL
9 MIL
7 MIL
DEC '07
SEPT '08
MARCH '09
JUNE '10
FOX NEWS
FOX NEWS
.COM
SOURCE: BLS
AMERICA'S NEWSROOM

Scale Distortions

"Double the axes, double the mischief"

(Quote from Gary Smith's *Standard Deviations*)

Graphic from Robert Reich's *Saving Capitalism*

http://www.thefunctionalart.com/2015/10/double-axes-double-mischief.html

Be Proportional

Trey Yingst @TreyYingst · May 11

Spotted: A map to be hung somewhere in the West Wing

Wash.
Mont.
N.D.
Ore.
Idaho
S.D.
Wyo.
Nev.
Utah
Colo.
Calif.
Ariz.
N.M.
Kan.
Okla.
Tex.
Minn.
Iowa
Mo.
Ark.
La.
Wis.
Ill.
Mich.
Ind.
Ohio
Ky.
Tenn.
Miss.
Ala.
Ga.
Fla.
S.C.
N.C.
Va.
W.Va.
Pa.
N.Y.
Vt.
Maine
N.H.
Mass.
R.I.
Conn.
Del.
Md.
D.C.
Alaska
Hawaii
Dem. lead, win
Rep. lead, win
Switched partyside last election

Wash.
Mont.
N.D.
Ore.
Idaho
S.D.
Wyo.
Nev.
Utah
Calif.
Ariz.
N.M.
Colo.
Kan.
Neb.
Iowa
Minn.
Wis.
Mich.
Ill.
Ind.
Ohio
Pa.
N.Y.
Vt.
Mass.
Conn.
N.J.
Del.
Md.
W.Va.
Va.
N.C.
S.C.
Ga.
Ala.
Miss.
La.
Ark.
Tex.
Okla.
Mo.
Tenn.
Ky.
Alaska
Hawaii

US Presidential Election 2016
Results mapped at county level showing the candidate with the largest vote share in each area
Overall result
Trump
62,979,636 votes (46.1%)
Clinton
65,844,610 votes (48.2%)
Other candidates
7,804,213 votes (5.7%)
Vote share
of candidate with most votes
0
50
70
90%
Trump
Clinton
Reference map
Gridded population cartogram
areas resized according to
the total number of people living there
Map updated January 2017
to reflect the final vote count
and outcome of the Electoral College
Map by Benjamin Hennig
www.viewsoftheworld.net

2016 Electoral Vote
ME 1 3
NH 4
VT 3
MA 11
CT 7
RI 4
NY 29
AK 3
WI 10
MN 10
MI 16
WA 8
3
1
IL 20
IA 6
IN 11
OH 18
PA 20
NJ 14
OR 7
ID 4
MT 3
WY 3
ND 3
SD 3
NE 5
WV 5
MD 10
DE 3
NV 5
UT 6
CO 9
KS 6
MO 10
KY 8
DC 3
VA 13
OK 7
AR 6
TN 11
NC 15
AZ 11
NM 5
TX 36
MS 6
AL 9
GA 16
SC 9
CA 55
LA 8
1
1
FL 29
1
HI 3
227 Clinton
304 Trump
Powell : 3
Spotted Eagle : 1
Paul : 1
Kasich : 1
Sanders : 1

Include Uncertainty

Note: The cone contains the probable path of the storm center but does not show the size of the storm. Hazardous conditions can occur outside of the cone.
MO
KY
WV
VA
TN
NC
AR
SC
GA
MS
AL
LA
FL
Bermuda
Bahamas
Cuba
Haiti
Dominican Rep.
Puerto Rico
Jamaica
Belize
2 AM Sun
2 AM Mon
2 AM Sat
2 AM Fri
2 AM Thu
8 AM Wed
35N
30N
25N
20N
95W
90W
85W
80W
75W
70W
65W
60W

What you show

What non-scientists are not aware of (cone is just 66% probability)

Alberto Cairo • University of Miami • www.thefunctionalart.com • Twitter: @albertocairo

What we could be showing instead

Plot all your data

Counties with the LOWEST kidney cancer death rates (1980-1989)

Counties with the LOWEST kidney cancer death rates (1980-1989)

Counties with the HIGHEST kidney cancer death rates (1980-1989)

0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
−0.0001
pct_mortality
0
100000
200000
300000
400000
500000
600000
700000
pop

2. Keep It Simple

Avoid Chartjunk

Extraneous visual elements that distract from the message

ongoing, Tim Brey

Volvo
Valiant
Toyota
Porsche
Pontiac
Merc
Mazda
Maserati
Lotus
Lincoln
Hornet
Honda
Ford
Fiat
Ferrari
Duster
Dodge
Datsun
Chrysler
Camaro
Cadillac
AMC
maker
0
5
10
15
20
25
30
1

Volvo
Valiant
Toyota
Porsche
Pontiac
Merc
Mazda
Maserati
Lotus
Lincoln
Hornet
Honda
Ford
Fiat
Ferrari
Duster
Dodge
Datsun
Chrysler
Camaro
Cadillac
AMC
maker
0
5
10
15
20
25
30
2

Volvo
Valiant
Toyota
Porsche
Pontiac
Merc
Mazda
Maserati
Lotus
Lincoln
Hornet
Honda
Ford
Fiat
Ferrari
Duster
Dodge
Datsun
Chrysler
Camaro
Cadillac
AMC
maker
0
5
10
15
20
25
30
3

Honda
Lotus
Fiat
Toyota
Porsche
Datsun
Volvo
Mazda
Hornet
Ferrari
Pontiac
Merc
Valiant
Ford
Dodge
AMC
Maserati
Chrysler
Duster
Camaro
Lincoln
Cadillac
30.4
30.4
29.85
27.7
26.0
22.8
21.4
21.0
20.05
19.7
19.2
19.0142857143
18.1
15.8
15.5
15.2
15.0
14.7
14.3
13.3
10.4
10.4
4

Don't!

matplotlib gallery

Excel Charts Blog

3. Use The Right Display

Chart Suggestions—A Thought-Starter

http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf

Comparisons

Bars vs. Lines

Zacks 1999

Proportions

Pie Charts

Stacked Bar Chart

S. Few

Stacked Area Chart

S. Few

Correlations

Scatterplots

http://xkcd.com/388/

London Cholera Epidemic

From Edward Tufte, Visual and Statistical Thinking

Don't!

matplot3d tutorial

Trends

Apple Inc. (AAPL) - NasdaqGS

Add to Portfolio | Like 6k

601.10 ↑15.53(2.65%)

4:00PM EDT | After Hours: **604.60** ↑3.50 (0.58%) 7:15PM EDT - Nasdaq Real Time Price

Enter name(s) or symbol(s) | GET CHART | COMPARE | EVENTS | TECHNICAL INDICATORS | CHART SETTINGS | RESET

Basic Chart | Full Screen | Print | Share | Send Feedback

Yahoo! Finance

35
30
25
20
15
10
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5

30
25
20
15
10
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5

Distributions

Histogram

ggplot2

Bin Width

binwidth = 0.1

binwidth = 0.01

ggplot2

Density Plots

https://
www.autodeskresearch.com/
publications/samestats

https://www.autodeskresearch.com/publications/samestats

GROUP

getting complex…

Faceting and Small Multiples

Use seaborn or
multiple plots in matplotlib

Small multiples

SPLOM

mpg
hp
wt
qsec
disp
35
30
25
20
15
10
350
300
250
200
150
100
50
5
4
3
2
22
20
18
16
14
400
300
200
100
20
40
0
200
400
2
4
6
15
20
25
0
250
500

Design Exercise

Hands-On 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

Come up with multiple visualizations.

Pen and Paper Only.

Pie

Side by side bar

Stacked bar, not very useful

Data Transposed Bar Chart

Difference Bar Chart

How do you feel about doing science?
Slopegraph
Level [%]
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Some interest, 68
Some interest, 44
OK, 40
Excited, 38
Kind of interested, 30
Kind of interested, 25
Excited, 19
OK, 14
Bored, 12
Bored, 11
Not great, 6
Not great, 5
Before
After

After the pilot program,

68%

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

Perceptual Effectiveness

Stephen's Power Law, 1961

	Nominal	Ordinal	Quantitative
Position	✔	✔	✔
Size	✔	✔	~
(Grey)Value	✔	✔	~
Texture	✔	~	✖
Color	✔	✖	✖
Orientation	✔	✖	✖
Shape	✔	✖	✖

✔ = Good
~ = OK
✖ = Bad

J. Bertin, 1967

Cleveland / McGill, 1984

J. Mackinlay, 1986

Heer / Bostock, 2010

How much longer?

How much steeper slope?

A

B

4x

How much larger area?

How much darker?

A

B

2x

How much bigger value?

A B

4x

C. Mulbrandon
VisualizingEconomics.com

Most Effective

VisualizingEconomics.com

Less Effective

VisualizingEconomics.com

Pie vs. Bar Charts

65% of the market is controlled by companies B and C

Least Effective

FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (*P*) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/*P* were made first at the 800-m resolution of the PRISM climate data. The mean values for the counties (shown) were then calculated by averaging the 800-m values within each county. Areas with fractions >1 are agricultural counties that either import surface water or mine deep groundwater.

Cliff Mass

Wash.
Mont.
N.D.
Ore.
Idaho
S.D.
Wyo.
Nev.
Utah
Colo.
Calif.
Ariz.
N.M.
Kan.
Okla.
Tex.
Minn.
Iowa
Mo.
Ark.
La.
Wis.
Ill.
Mich.
Ind.
Ohio
Ky.
Tenn.
Miss.
Ala.
Ga.
Fla.
S.C.
N.C.
Va.
W.Va.
Pa.
N.Y.
Vt.
Maine
N.H.
Mass.
R.I.
Conn.
Del.
Md.
D.C.
Alaska
Hawaii
Dem. lead, win
Rep. lead, win
Switched partyside last election

Wash.
Mont.
N.D.
Ore.
Idaho
S.D.
Wyo.
Nev.
Utah
Calif.
Ariz.
N.M.
Colo.
Kan.
Neb.
Iowa
Minn.
Wis.
Mich.
Ill.
Ind.
Ohio
Pa.
N.Y.
Vt.
Mass.
Conn.
N.J.
Del.
Md.
W.Va.
Va.
N.C.
S.C.
Ga.
Ala.
Miss.
La.
Ark.
Tex.
Okla.
Mo.
Tenn.
Ky.
Alaska
Hawaii

US Presidential Election 2016
Results mapped at county level showing the candidate with the largest vote share in each area
Overall result
Trump
62,979,636 votes (46.1%)
Clinton
65,844,610 votes (48.2%)
Other candidates
7,804,213 votes (5.7%)
Vote share
of candidate with most votes
0
50
70
90%
Trump
Clinton
Reference map
Gridded population cartogram
areas resized according to
the total number of people living there
Map updated January 2017
to reflect the final vote count
and outcome of the Electoral College
Map by Benjamin Hennig
www.viewsoftheworld.net

2016 Electoral Vote
ME 1 3
NH 4
VT 3
MA 11
CT 7
RI 4
NY 29
AK 3
WI 10
MN 10
MI 16
WA 8
3
1
IL 20
IA 6
IN 11
OH 18
PA 20
NJ 14
OR 7
ID 4
MT 3
WY 3
ND 3
SD 3
NE 5
WV 5
MD 10
DE 3
NV 5
UT 6
CO 9
KS 6
MO 10
KY 8
DC 3
VA 13
OK 7
AR 6
TN 11
NC 15
AZ 11
NM 5
TX 36
MS 6
AL 9
GA 16
SC 9
CA 55
LA 8
1
1
FL 29
1
HI 3
227 Clinton
304 Trump
Powell : 3
Spotted Eagle : 1
Paul : 1
Kasich : 1
Sanders : 1

Data visualization and visual encoding

Length or height

Position

Area

Angle/area

Line weight

Hue and shade

Figures represented
in all these graphics:
22%, 25%, 34%, 29%, 32%

Alberto Cairo • University of Miami • www.thefunctionalart.com • Twitter: @albertocairo

4. Use Color Strategically

Colors for Categories

Do not use more than 5-8 colors at once

Ware, "Information Visualization"

Colors for Ordinal Data

Vary luminance and saturation

Zeilis et al, 2009, "Escaping RGBland: Selecting Colors for Statistical Graphics"

Colors for Quantitative Data

Hue
(Rainbow)

Luminance

Luminance
& Hue

Rogowitz and Treinish, Why should engineers and scientists be worried about color?

Rainbow Colormap

Rainbow Colormap

Perceptually nonlinear

R. Simmon

Avoid Rainbow Colors!

matplotlib gallery

Gray

Colormap evaluation: gray
The colormap in its glory
Black-and-white printed
Sample images
Moderate deuter
Moderate deuteranomaly
Moderate protanomaly
Complete deuteranopia
Complete protanopia
Toggle gamut

Color Blindness

Protanope Deuteranope Tritanope

Red / green deficiencies

Blue / Yellow deficiency

Based on slide from Stone

Color Blindness

Normal Protanope Deuteranope Lightness

Based on slide from Stone

Viridis

Colormap evaluation: option_d.py
The colormap in its glory
Black-and-white printed
Sample images
Moderate deuter.
Perceptual deltas (total: 141.79)
Perceptual lightness deltas (total: 73.63)
Moderate deuteranomaly
Moderate protanomaly
Complete deuteranopia
Complete protanopia
Toggle gamut

Color Brewer

Nominal

Ordinal

Cynthia Brewer, Color Use Guidelines for Data Representation

number of data classes on your map

3 learn more >

the nature of your data

sequential learn more >

pick a color scheme: GnBu

multihue

single hue

(optional) only show schemes that are:

- [x] colorblind safe
- [ ] print friendly
- [ ] photocopy-able

learn more >

pick a color system

224, 243, 219

168, 221, 181

67, 162, 202

(•) RGB ( ) CMYK ( ) HEX

adjust map context

- [ ] roads
- [ ] cities
- [x] borders

select a background

(•) solid color

( ) terrain

color transparency

learn more >

how to use | updates | credits

COLORBREWER 2.0

color advice for cartography

EXPORT YOUR COLORS >>

SCORE CARD

© Cynthia Brewer, Mark Harrower and The Pennsylvania State University

Support

Back to ColorBrewer 1.0

axm

Diverging Palette for Quantitative or Ordinal

Sequential Palette for Densities

Effective Visualizations

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

Further Reading

Edward Tufte

Stephen Few

Show Me the Numbers
Designing Tables and Graphs to Enlighten
Stephen Few

INFORMATION
DASHBOARD
DESIGN
The Effective Visual Communication of Data
Stephen Few
O'REILLY

Now you see it
Stephen Few

2012

2016

I've always believed in the power of data visualization (the representation of information by means of charts, diagrams, maps, etc.) to enable understanding