

[COGS 9] Discussion Reading 3, Visualizations, Matplotlib

Reading Quiz 3 due on 19th July (Wed)

Assignment 1 due on 20th July (Thu)

Final Project discussion on 20th July (Thu)



Why Tidy Data?

- Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table.



What is dataset?

- A dataset is a collection of values, usually either numbers (if quantitative) or strings (if qualitative)
- Values are organized in two ways. Every value belongs to a variable and an observation.

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Problems with Messy Dataset

Column headers are values, not variable names.

Multiple variables are stored in one column.

Variables are stored in both rows and columns.

Multiple types of observational units are stored in the same table. ❓

A single observational unit is stored in multiple tables.

Column headers are values, not variable names

- Definition: Melting
 - Turning columns into rows
 - Parametrizing a list of columns that are already variables and covert the other columns into variables containing repeated column headings and the concatenated data values from the previous separate columns

row	a	b	c
A	1	4	7
B	2	5	8
C	3	6	9

(a) Raw data

row	column	value
A	a	1
B	a	2
C	a	3
A	b	4
B	b	5
C	b	6
A	c	7
B	c	8
C	c	9

(b) Molten data

Multiple variables stored in one column

After melting, we need to split the column column into columns each containing one kind of data.

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

(a) Molten data

country	year	sex	age	cases
AD	2000	m	0–14	0
AD	2000	m	15–24	0
AD	2000	m	25–34	1
AD	2000	m	35–44	0
AD	2000	m	45–54	0
AD	2000	m	55–64	0
AD	2000	m	65+	0
AE	2000	m	0–14	2
AE	2000	m	15–24	4
AE	2000	m	25–34	4
AE	2000	m	35–44	6
AE	2000	m	45–54	5
AE	2000	m	55–64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

(b) Tidy data

Variables are stored in both rows and columns

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—

id	date	element	value
MX17004	2010-01-30	tmax	27.8
MX17004	2010-01-30	tmin	14.5
MX17004	2010-02-02	tmax	27.3
MX17004	2010-02-02	tmin	14.4
MX17004	2010-02-03	tmax	24.1
MX17004	2010-02-03	tmin	14.4
MX17004	2010-02-11	tmax	29.7
MX17004	2010-02-11	tmin	13.4
MX17004	2010-02-23	tmax	29.9
MX17004	2010-02-23	tmin	10.7

(a) Molten data

id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

(b) Tidy data

Multiple types in one table



Datasets often involve values collected at multiple levels, on different types of observational units.



During tidying, each type of observational unit should be stored in its own table.



This is closely related to the idea of database normalization, where each fact is expressed in only one place. (could lead to potential inconsistencies within the df)

id	artist	track	time	id	date	rank
1	2 Pac	Baby Don't Cry	4:22	1	2000-02-26	87
2	2Ge+her	The Hardest Part Of ...	3:15	1	2000-03-04	82
3	3 Doors Down	Kryptonite	3:53	1	2000-03-11	72
4	3 Doors Down	Loser	4:24	1	2000-03-18	77
5	504 Boyz	Wobble Wobble	3:35	1	2000-03-25	87
6	98~0	Give Me Just One Nig...	3:24	1	2000-04-01	94
7	A*Teens	Dancing Queen	3:44	1	2000-04-08	99
8	Aaliyah	I Don't Wanna	4:15	2	2000-09-02	91
9	Aaliyah	Try Again	4:03	2	2000-09-09	87
10	Adams, Yolanda	Open My Heart	5:30	2	2000-09-16	92
11	Adkins, Trace	More	3:05	3	2000-04-08	81
12	Aguilera, Christina	Come On Over Baby	3:38	3	2000-04-15	70
13	Aguilera, Christina	I Turn To You	4:00	3	2000-04-22	68
14	Aguilera, Christina	What A Girl Wants	3:18	3	2000-04-29	67
15	Alice DeeJay	Better Off Alone	6:50	3	2000-05-06	66



Tidy Tools

- Tidying data makes it easier to maintain and do analysis with.
- Manipulation functions:
 - Filter: subsetting or removing observations based on some condition.
 - Transform: adding or modifying variables. These modifications can involve either a single variable (e.g., log-transformation), or multiple variables (e.g., computing density from weight and volume).
 - Aggregate: collapsing multiple values into a single value (e.g., by summing or taking means).
 - Sort: changing the order of observations.

Data organization in spreadsheets

Be consistent

Use	consistent codes for categorical variables
Use	a consistent fixed code for any missing values
Use	consistent variable names
Use	consistent subject identifiers
Use	a consistent data layout in multiple files
Use	consistent file names
Use	a consistent format for all dates
Use	consistent phrases in your notes
Be	careful about extra spaces within cells

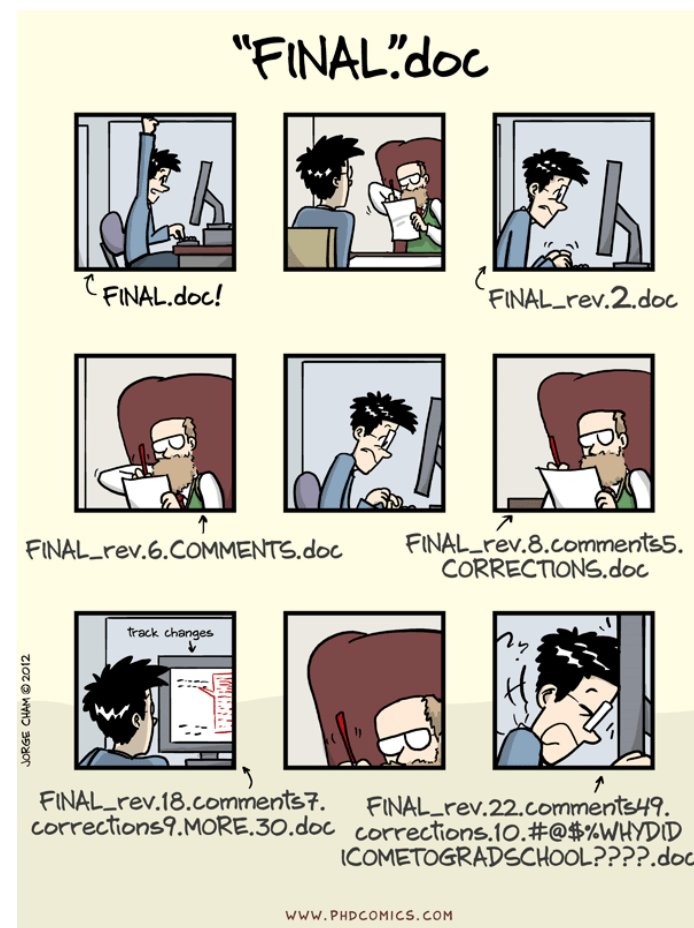
- Male, M, male, m... (Stick to one of them!)
- N/A, NaN, Null ☒ -999, 999 ✗
- grades_wk10, wk10Grades...
- studentA, aStudent, a...
- -
- "quiz_010123.csv", "010123q.csv"
- YYYY-MM-DD
- -
- "Male" ≠ " Male"



Choose good names for things

Table 1: Examples of good and bad variable names.

good name	good alternative	avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
Observation_01	first_observation	1st Obs.



Write dates as YYYY-MM-DD and No empty cells

- ISO 8601 standard


PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS ***THE*** CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013.II.27. 27½-13 2013.158904109
MMXIII-II-XXVII MMXIII ^{LVII}/_{CCCLXV} 1330300800
 $((3+3) \times (111+1) - 1) \times 3 / 3 - 1 / 3^3$ 2013 
10/11011/1101 02/27/20/13 $\begin{smallmatrix} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ 5 & 6 & 7 & 8 \end{smallmatrix}$

	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102		95.3
4	103	2015-06-18	97.5
5	104		117.0
6	105		108.0
7	106	2015-06-20	149.0
8	107		169.4

Put just one thing in a cell

	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172
10	A	normal	5	1	334
11	A	normal	5	2	354
12	B	normal	5	1	514
13	B	normal	5	2	611
14	A	mutant	5	1	451
15	A	mutant	5	2	474
16	B	mutant	5	1	412
17	B	mutant	5	2	447

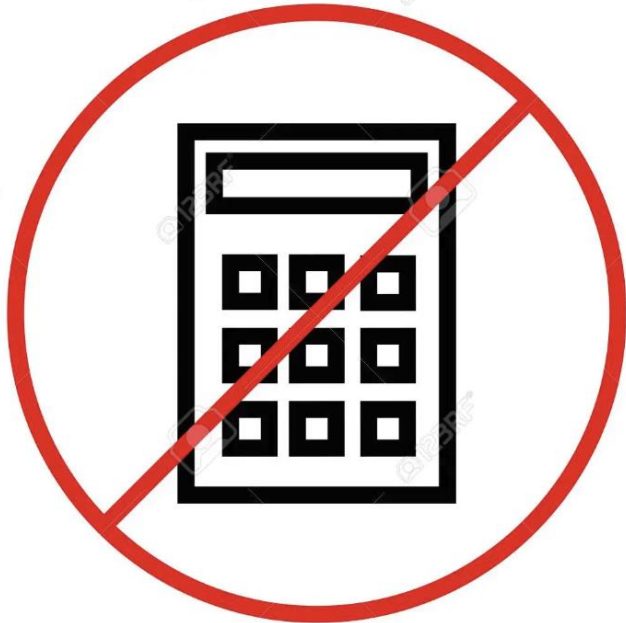
Figure 3: A tidy version of the data in Figure 2B.

Create a data dictionary

	A	B	C	D
1	name	plot_name	group	description
2	mouse	Mouse	demographic	Animal identifier
3	sex	Sex	demographic	Male (M) or Female (F)
4	sac_date	Date of sac	demographic	Date mouse was sacrificed
5	partial_inflation	Partial inflation	clinical	Indicates if mouse showed partial pancreatic inflation
6	coat_color	Coat color	demographic	Coat color, by visual inspection
7	crumblers	Crumblers	clinical	Indicates if mouse stored food in their bedding
8	diet_days	Days on diet	clinical	Number of days on high-fat diet

No calculations in
the raw data files

Make Backups



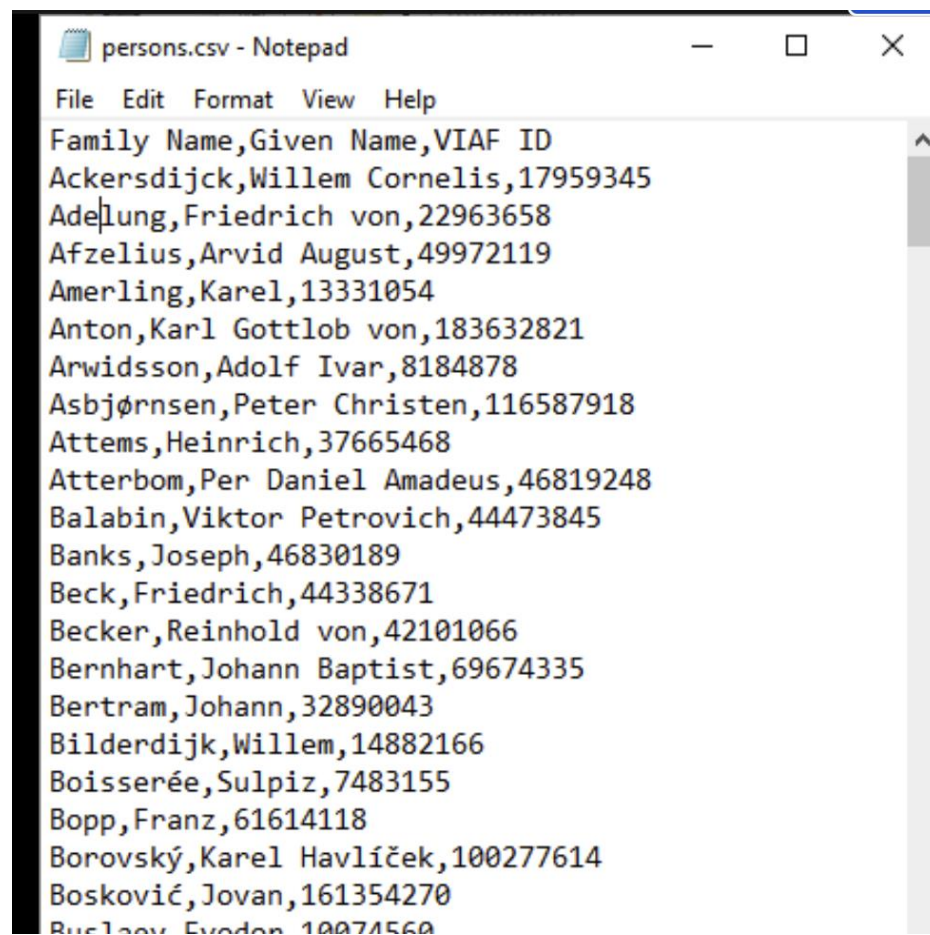
Font color/Highlighting

- Nice visually, but hard for later analysis
- Better encode highlight info in another column

	A	B	C	D	E	F	G	H
1		Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus
2	Mass (10^{24} kg)	0.33	4.87	5.97	0.642	1898	568	86.8
3	Diameter (km)	4879	12,104	12,756	6792	142,984	120,536	51,118
4	Density (kg/m^3)	5427	5243	5514	3933	1326	687	1271
5	Gravity (m/s^2)	3.7	8.9	9.8	3.7	23.1	9	8.7
6	Escape Velocity (km/s)	4.3	10.4	11.2	5	59.5	35.5	21.3
7	Rotation Period (hours)	1407.6	-5832.5	23.9	24.6	9.9	10.7	-17.2
8	Length of Day (hours)	4222.6	2802	24	24.7	9.9	10.7	17.2
9	Distance from Sun (10^6 km)	57.9	108.2	149.6	227.9	778.6	1433.5	2872.5
10	Perihelion (10^6 km)	46	107.5	147.1	206.6	740.5	1352.6	2741.3
11	Aphelion (10^6 km)	69.8	108.9	152.1	249.2	816.6	1514.5	3003.6
12	Orbital Period (days)	88	224.7	365.2	687	4331	10,747	30,589
13	Orbital Velocity (km/s)	47.4	35	29.8	24.1	13.1	9.7	6.8
14	Orbital Inclination	7	3.4	0	1.9	1.3	2.5	0.8
15	Orbital Eccentricity	0.205	0.007	0.017	0.094	0.049	0.057	0.046

Don't use them!

Save the data in plain text files



```
persons.csv - Notepad
File Edit Format View Help
Family Name,Given Name,VIAF ID
Ackersdijck,Willem Cornelis,17959345
Adelung,Friedrich von,22963658
Afzelius,Arvid August,49972119
Amerling,Karel,13331054
Anton,Karl Gottlob von,183632821
Arwidsson,Adolf Ivar,8184878
Asbjørnsen,Peter Christen,116587918
Attems,Heinrich,37665468
Atterbom,Per Daniel Amadeus,46819248
Balabin,Viktor Petrovich,44473845
Banks,Joseph,46830189
Beck,Friedrich,44338671
Becker,Reinhold von,42101066
Bernhart,Johann Baptist,69674335
Bertram,Johann,32890043
Bilderdijk,Willem,14882166
Boisserée,Sulpiz,7483155
Bopp,Franz,61614118
Borovský,Karel Havlíček,100277614
Bosković,Jovan,161354270
Buslow,Eugen,10074560
```

Make it a rectangle

- First row should contain variable names
- Don't use more than one row for variable names

A

	A	B	C	D	E	F
1						
2		101	102	103	104	105
3	sex	Male	Female	Male	Male	Male
4						
5		101	102	103	104	105
6	glucose	134.1	120.0	124.8	83.1	105.2
7						
8		101	102	103	104	105
9	insulin	0.60	1.18	1.23	1.16	0.73

C

	A	B	C	D	E	F	G
1							
2	Date	11/3/14					
3	Days on diet	126					
4	Mouse #	43					
5	sex	f					
6	experiment		values			mean	SD
7	control		0.186	0.191	1.081	0.49	0.52
8	treatment A		7.414	1.468	2.254	3.71	3.23
9	treatment B		9.811	9.259	11.296	10.12	1.05
10							
11	fold change		values			mean	SD
12	treatment A		15.26	3.02	4.64	7.64	6.65
13	treatment B		20.19	19.05	23.24	20.83	2.17

B

	A	B	C	D	E	F	G
1	1MIN						
2			Normal			Mutant	
3	B6	146.6	138.6	155.6	166	179.3	186.9
4	BTBR	245.7	240	243.1	177.8	171.6	188.1
5							
6	5MIN						
7			Normal			Mutant	
8	B6	333.6	353.6	408.8	450.6	474.4	423.8
9	BTBR	514.4	610.6	597.9	412.1	447.4	446.5

D

	A	B	C	D	E	F
1		GTT date	GTT weight	time	glucose mg/dl	insulin ng/ml
2	321	2/9/15	24.5	0	99.2	lo off curve
3				5	349.3	0.205
4				15	286.1	0.129
5				30	312	0.175
6				60	99.9	0.122
7				120	217.9	lo off curve
8	322	2/9/15	18.9	0	185.8	0.251
9				5	297.4	2.228
10				15	439	2.078
11				30	362.3	0.775
12				60	232.7	0.5
13				120	260.7	0.523
14	323	2/9/15	24.7	0	198.5	0.151
15				5	530.6	off curve lo

	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

Programming for Data Science

Why and How to Program?

- Tips on how to learn to code and resources

Different roles in the industry

- Software Engineer, Software Developer, Computer Scientist
- Data Scientist, Machine Learning Engineer, Data Analyst

R (in Academia), Python, SQL

- Imperative vs Declarative

Version Control (Git) -> Github

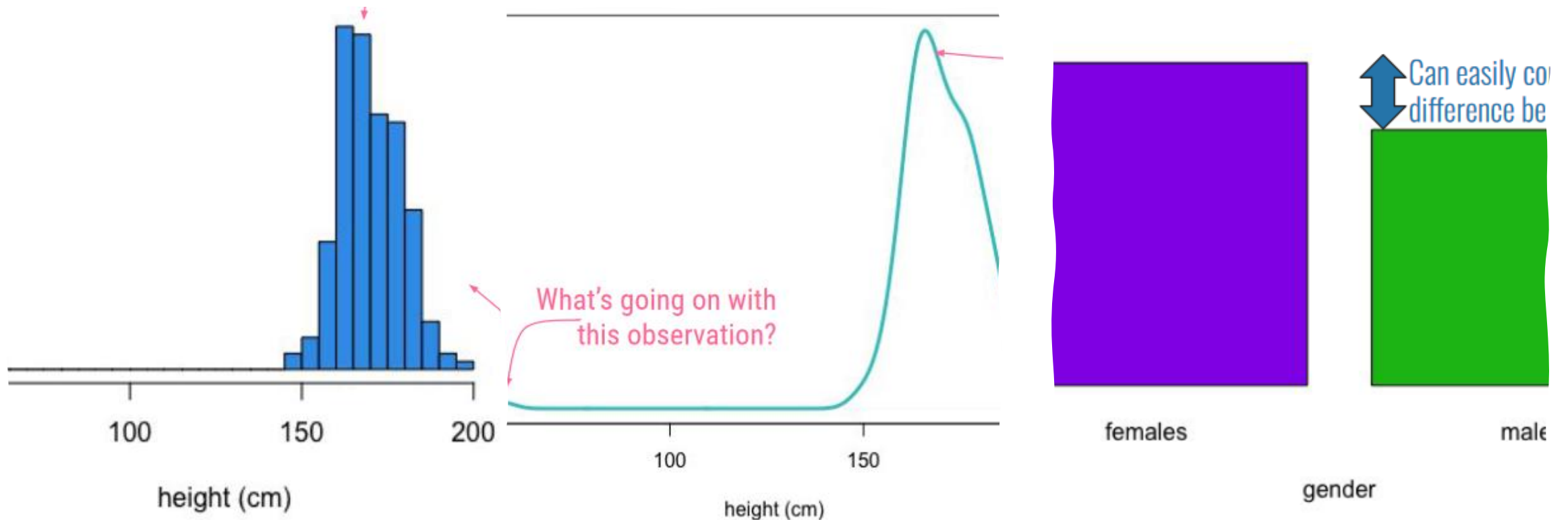
- GOLD MINE - <https://education.github.com/pack/offers>

Python Data Stack

- Anaconda, Miniconda, Numpy, Pandas, Matplotlib, Seaborn, Scipy, SKLearn, PyTorch

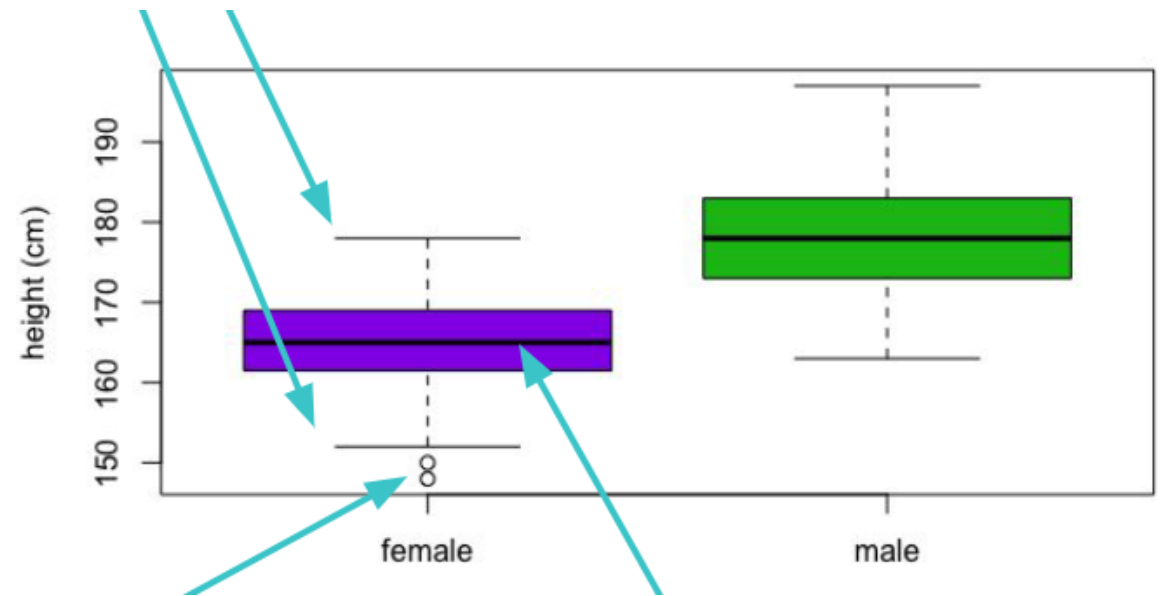
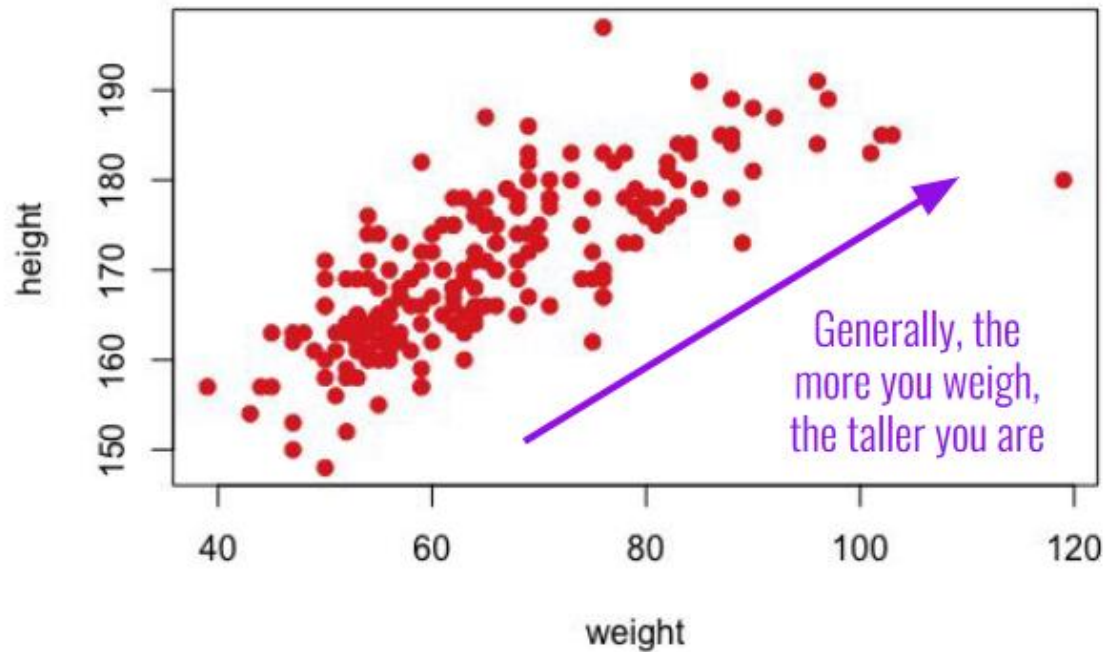
Data Visualization

What's the difference between a histogram, densityplot and barplot?



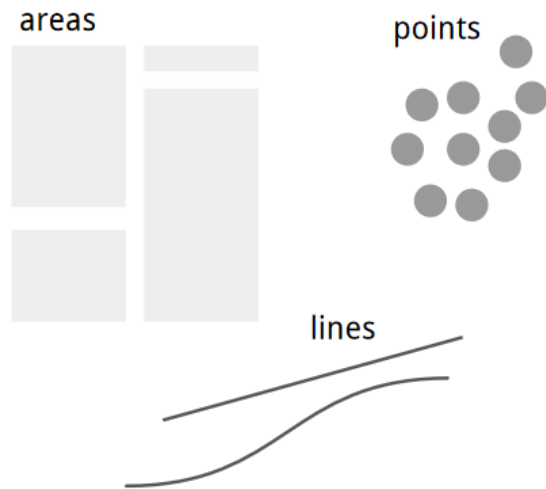
Data Visualization

When to use a scatterplot and a boxplot?

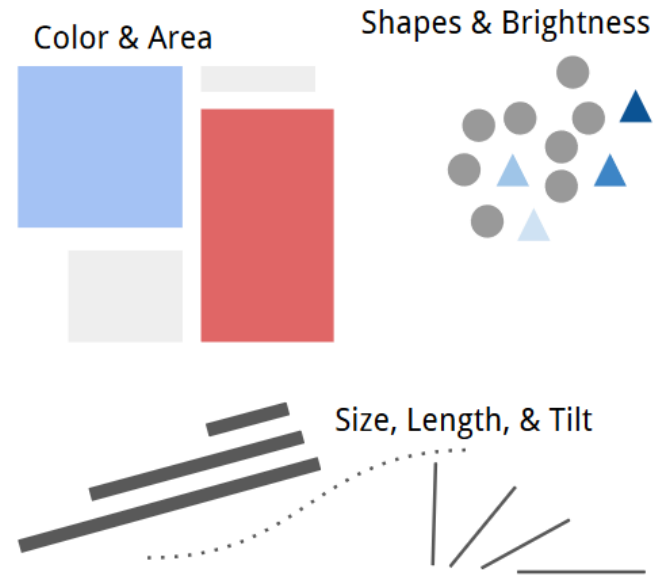


Data Visualization

Marks and channels



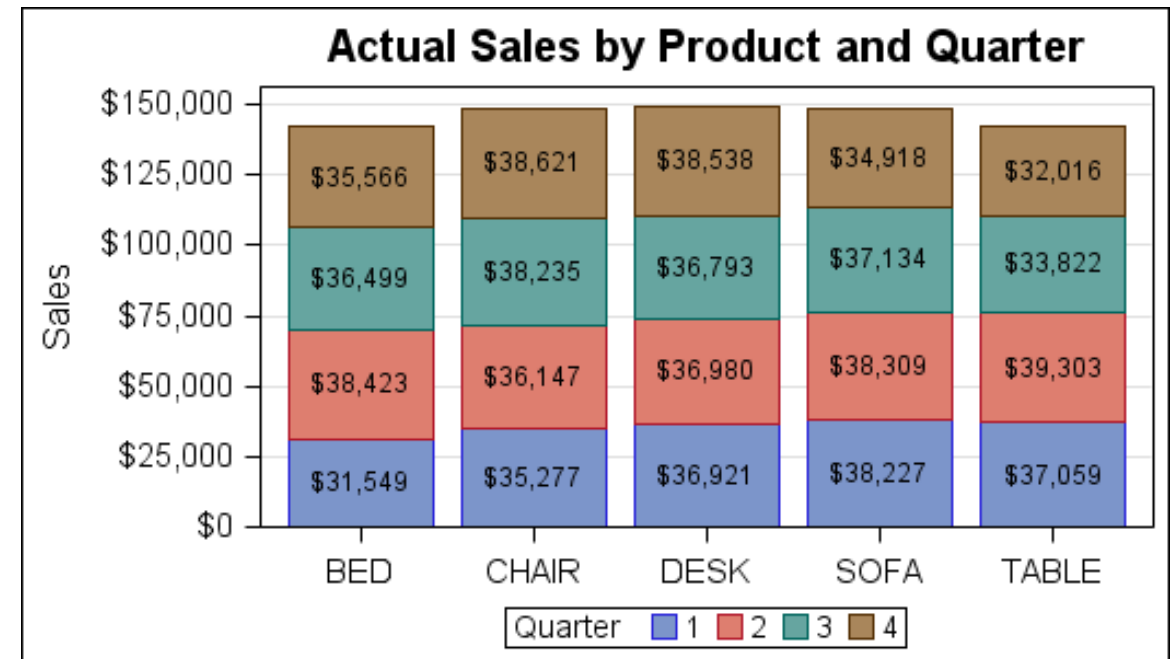
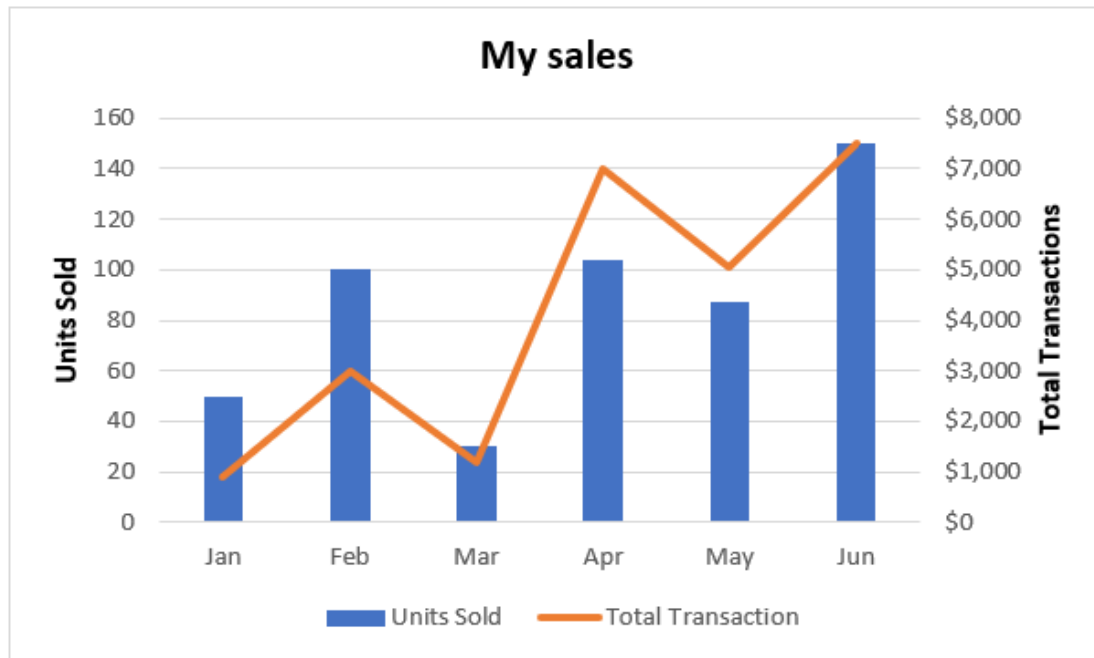
Marks: Geometric Primitives



Channel: way to control the appearance of a Mark

Data Visualization

- Express: The visual shall express all of, and only, the information provided by the data's attributes. The visual should not add anything to/remove from the data.
- Effect: How important an attribute is, must match the salience of the channel. Greater importance = greater salience, or more noticeable



Data Visualization

Checklist when creating graphs

- Consideration for Colour-blindness
- Label the axes
- Ensure that the data is correct
- Ensure that the graphic represents the data
- Make the comparison easy on readers
- Ensure that the y-axis starts at 0 (What about x-axis?)
- Choose best visual
- Keep it Simple Stupid

Data Visualization

Checklist when creating tables

- Have a top to bottom comparison
- Logical row ordering
- Logical column ordering
- Limit number of rows and columns
- Informative headers
- Fix significant digits
- Format table

Descriptive Analysis

- Suppress some of the truth so that humans can understand easily
- Size, shape, missingness, central tendency, variability
- Size: Number of variables and observations
- Shape: Distribution of the variables (Uniform, bimodal, Normal/Gaussian/Bell-shaped, left & right skewed, random)
- Missingness: How much data is missing?
- Central Tendency: Mean, median, mode
- Variability: Variance, Standard Deviation, Range

Matplotlib Demo

<https://matplotlib.org/stable/tutorials/index.html>

<https://matplotlib.org/stable/gallery/index.html>

