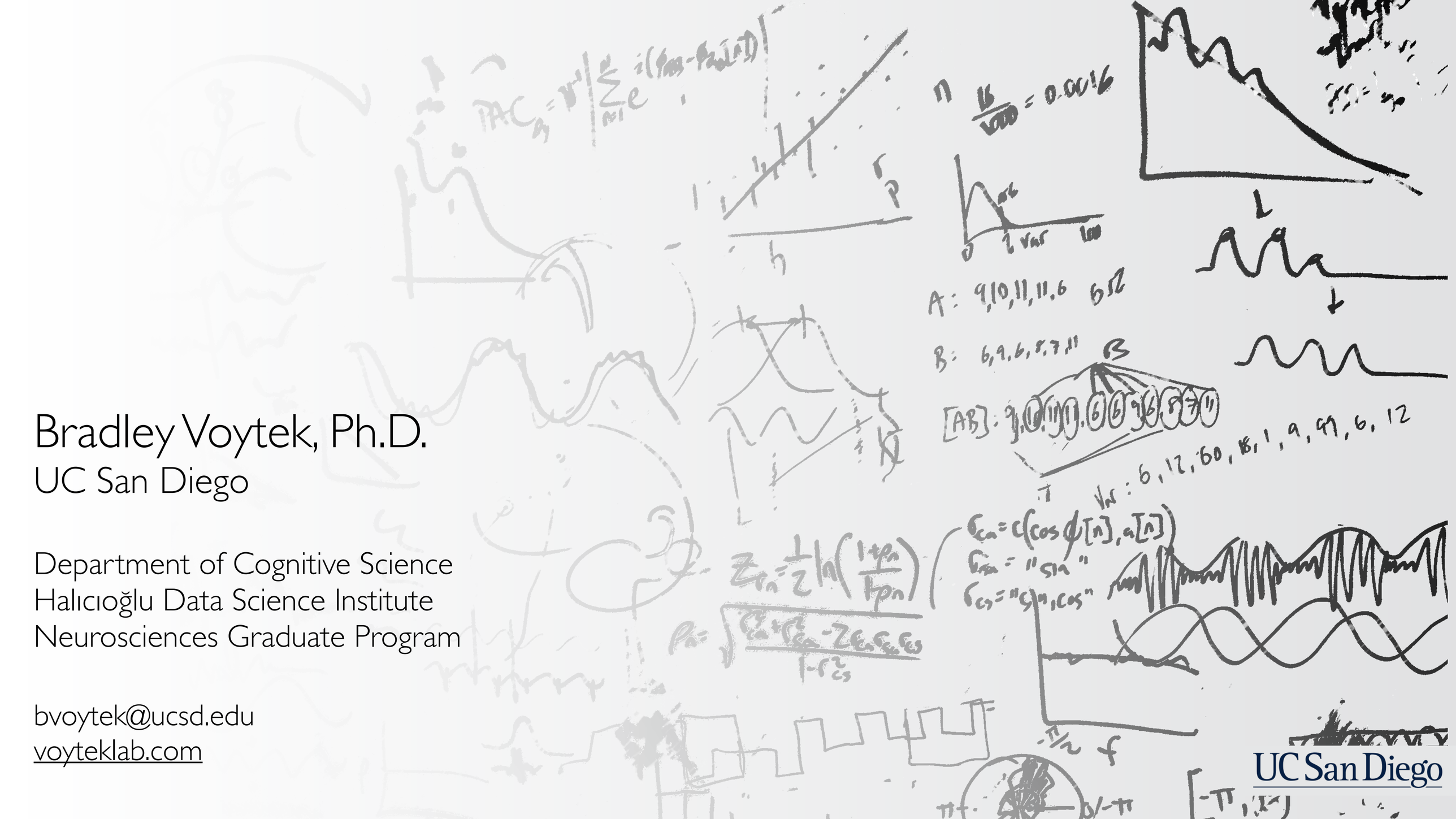
Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Halıcıoğlu Data Science Institute
Neurosciences Graduate Program

bvoytek@ucsd.edu
voyteklab.com



UC San Diego

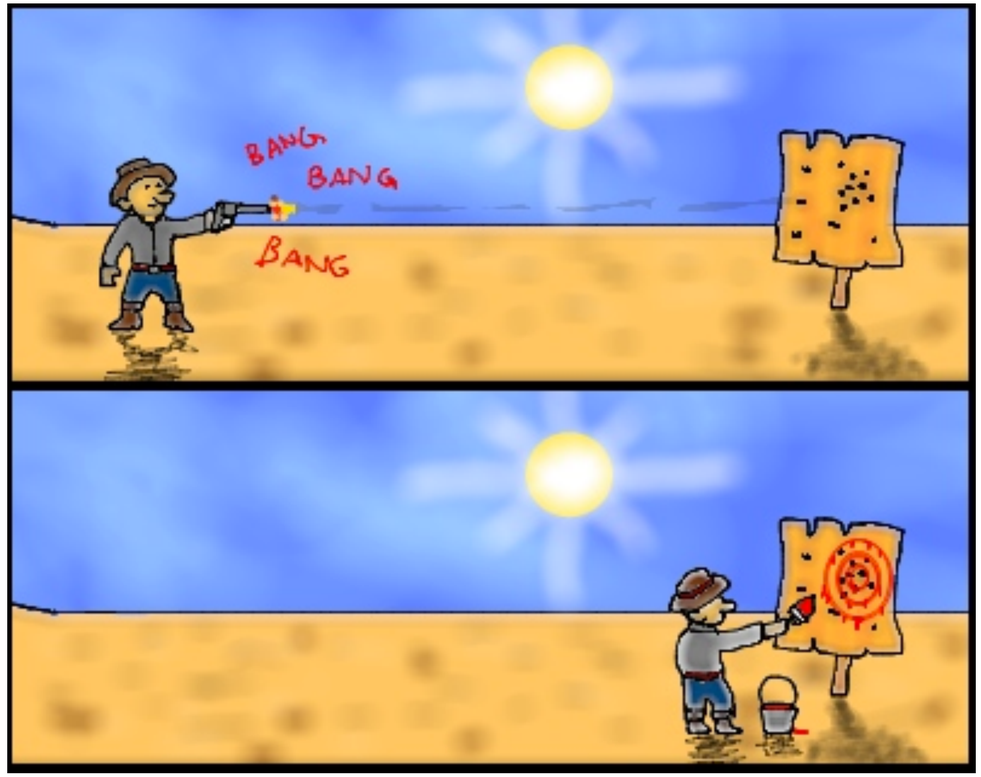
A3 is released!

COGS 9
Introduction to Data Science

Statistical inference and ML
(stray thoughts)

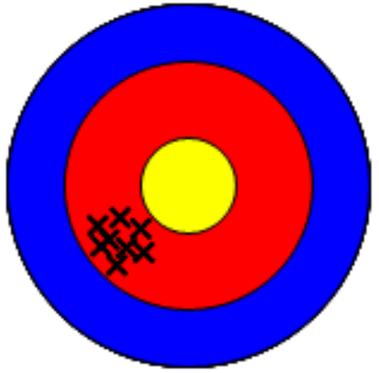
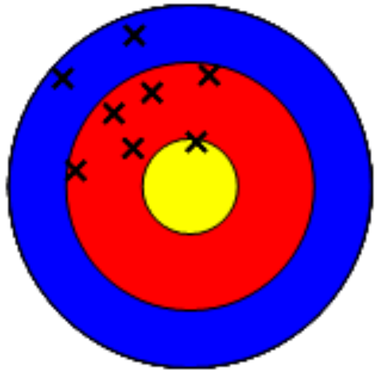
UC San Diego

Texas sharpshooter fallacy



The fallacy is characterized by a lack of a specific hypothesis prior to the gathering of data, and the formulation of a hypothesis only after data have already been gathered and examined.

Accuracy vs. Precision

	Accurate	Inaccurate (systematic error)
Precise		
Imprecise (reproducibility error)		

Fermi Estimation

Paint the Earth

Has humanity produced enough paint to cover the entire land area of the Earth?

—Josh (Bolton, MA)

Source: *XKCD*, What If?

Fermi Estimation

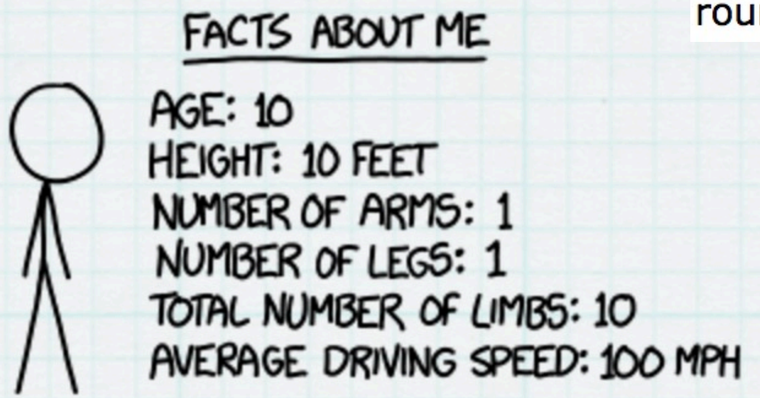
This answer is pretty straightforward. We can look up the size of the world's paint industry, extrapolate backward to figure out the total amount of paint produced. We'd also need to make some assumptions about how we're painting the ground. Note: When we get to the Sahara desert, I recommend not using a brush.

Source: *XKCD*, What If?

Fermi Estimation

But first, let's think about different ways we might come up with a guess for what the answer will be. In this kind of thinking—often called **Fermi estimation**—all that matters is getting in the right ballpark; that is, the answer should have about the right number of digits. In Fermi estimation, you can round[1] all your answers to the nearest order of magnitude:

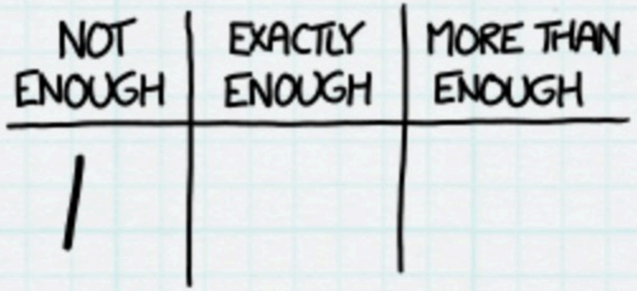
Using the formula $\text{Fermi}(x) = 10^{\text{round}(log_{10}x)}$, meaning that 3 rounds to 1 and 4 rounds to 10.



Source: *XKCD*, What If?

Fermi Estimation

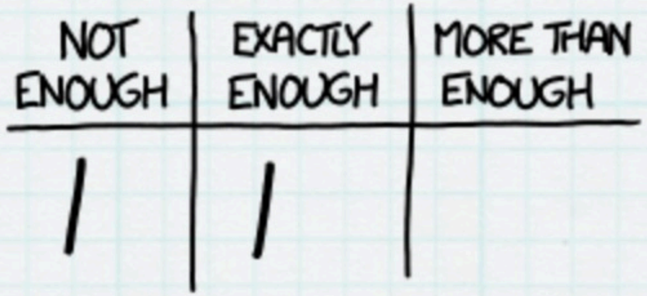
Let's suppose that, on average, everyone in the world is responsible for the existence of two rooms, and they're both painted. My living room has about 50 square meters of paintable area, and two of those would be 100 square meters. 7.15 billion people times 100 square meters per person is a little under a trillion square meters—an area smaller than Egypt.



Source: *XKCD*, What If?

Fermi Estimation

Let's make a wild guess that, on average, one person out of every thousand spends their working life painting things. If I assume it would take me three hours to paint the room I'm in, [2] and 100 billion people have ever lived, and each of them spent 30 years painting things for 8 hours a day, we come up with 150 trillion square meters ... just about exactly the land area of the Earth.



Source: *XKCD*, What If?

Fermi Estimation

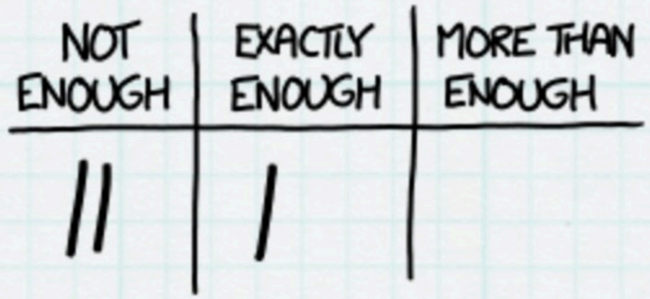
How much paint does it take to paint a house? I'm not enough of an adult to have any idea, so let's take another Fermi guess.

Based on my impressions from walking down the aisles, home improvement stores stock about as many light bulbs as cans of paint. A normal house might have about 20 light bulbs, so let's assume a house needs about 20 gallons of paint.[3] Sure, that sounds about right.

Source: *XKCD*, What If?

Fermi Estimation

The average US home costs about $200,000. Assuming each gallon of paint covers about 300 square feet, that's a square meter of paint per $300 of real estate. I vaguely remember that the world's real estate has a combined value of something like $100 trillion,[4] which suggests there's about 300 billion square meters of paint on the world's real estate. That's about one New Mexico.



Source: *XKCD*, What If?

Fermi Estimation

Of course, both of the building-related guesses could be overestimates (lots of buildings are not painted) or underestimates (lots of things that are not buildings[5] are painted) But from these wild Fermi estimates, my guess would be that there probably isn't enough paint to cover all the land.

So, how did Fermi do?

Source: *XKCD*, What If?

Fermi Estimation

According to the report **The State of the Global Coatings Industry**, the world produced 34 billion liters of paints and coatings in 2012.

There's a neat trick that can help us here. If some quantity—say, the world economy—has been growing for a while at an annual rate of **n**—say, 3% (0.03)—then the most recent year's share of the whole total so far is $1 - \frac{1}{1+n}$, and the whole total so far is the most recent year's amount times $1 + \frac{1}{n}$.

Source: *XKCD*, What If?

Fermi Estimation

If we assume paint production has, in recent decades, followed the economy and grown at about 3% per year, that means the total amount of paint produced equals the current yearly production times 34.[6] That comes out to a little over a trillion liters of paint. At 30 square meters per gallon,[7] that's enough to cover 9 trillion square meters—about the area of the United States.

So the answer is no; there's not enough paint to cover the Earth's land, and—at this rate—probably won't be enough until the year 2100.

Source: *XKCD*, What If?

COGS 9
Introduction to Data Science

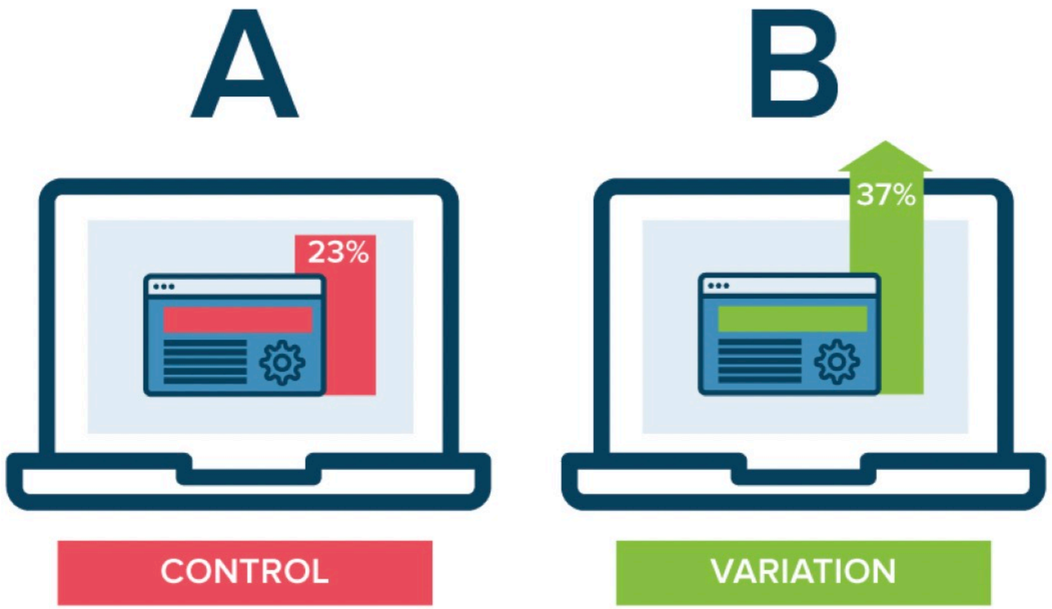
A/B testing

UC San Diego

Today's Learning Objective

- *Describe A/B testing and examples of when it would be used*
- *Identify issues in A/B testing and explain how to avoid them*

A/B testing



A/B Testing : Designing and running an experiment to compare two versions (typically, a web page or an app), to determine which is "better"

Classic Example: If you change the color of a button, and measure the click-rate, which is higher?

Source: https://www.optimizely.com/optimization-glossary/ab-testing/

Guidelines for A/B testing

1. Choose one key metric for testing

a. You can/should *monitor* multiple metrics, but only one should be your metric for measuring effect
b. Generally, choose a proportion metric (%/proportion that click is better than revenue)
c. Why? False positives

	Real Difference	No Real Difference
Test Positive	True Positive	False Positive (controlled by the significance level)
Test Negative	False Negative (controlled by the power level)	True Negative

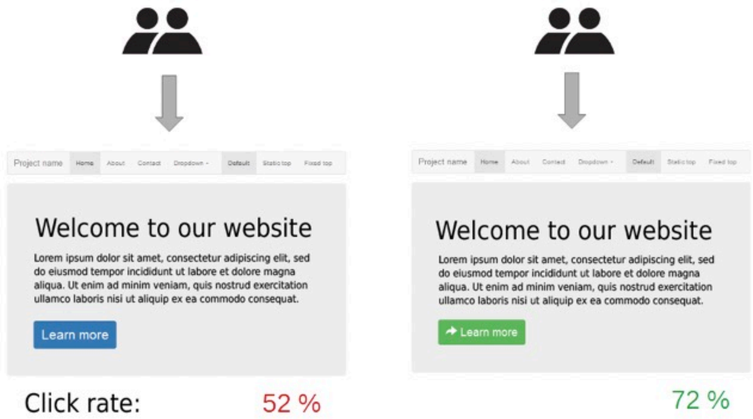
https://docs.adobe.com/content/help/en/target/using/activities/abtest/common-ab-testing-pitfalls.html

Source: https://hookedondata.org/guidelines-for-ab-testing/

Guidelines for A/B testing

2. Design Experiment & Run for Length of Time Planned

a. *Power Calculations* help determine how long your experiment should run
b. Determine your null and alternative hypotheses.
c. If you plan for two weeks and peak at results after a week, and see awesome results (CI is *far* from zero; p << 0.05)….keep running for your second week
d. (If you don't, this is an example of p-hacking. Don't be a p-hacker.)
e. Issue? False Positives



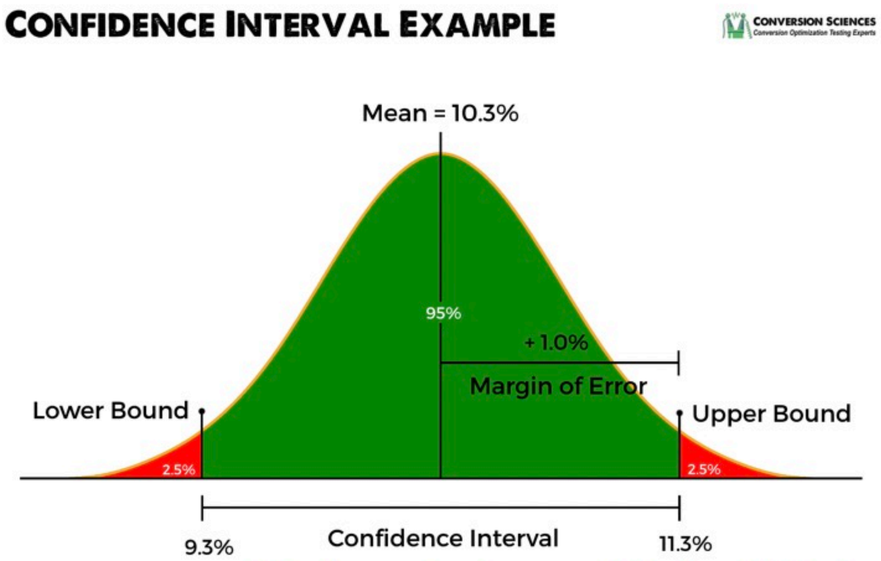
https://en.wikipedia.org/wiki/A/B_testing

Source: https://hookedondata.org/guidelines-for-ab-testing/

Guidelines for A/B testing

3. Confidence Intervals more important than p-values

a. Note: Confidence intervals and p-values are related
b. Confidence intervals are a better indicator of *how* much of a difference there is between A and B
c. Issue? Large samples can result in small p-values, even for very small differences



Source: https://hookedondata.org/guidelines-for-ab-testing/

UC San Diego

Guidelines for A/B testing

4. Don't look at all the possible subgroups

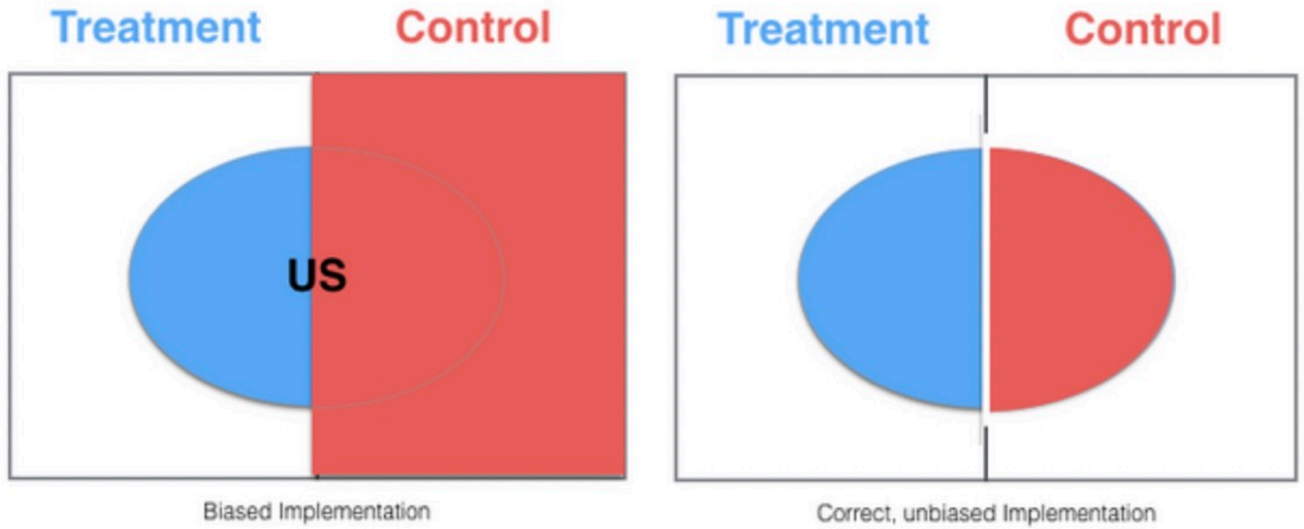
a. US visitors? New visitors? Visitors on Sunday?
b. Issues? False positives and multiple testing
c. If you want to look at a subgroup, design experiment to look at that before the experiment is run

Source: https://hookedondata.org/guidelines-for-ab-testing/

Guidelines for A/B testing

5. Look for "bucketing skew"

a. You planned for A and B variants to have the same number of participants….did that happen? You should check!
b. If you see skew, identify why? B/c you have a bug…
    i. i.e. Did B take longer to load and people w/ slow connections left?
    ii. Once found, fix, and re-run experiment



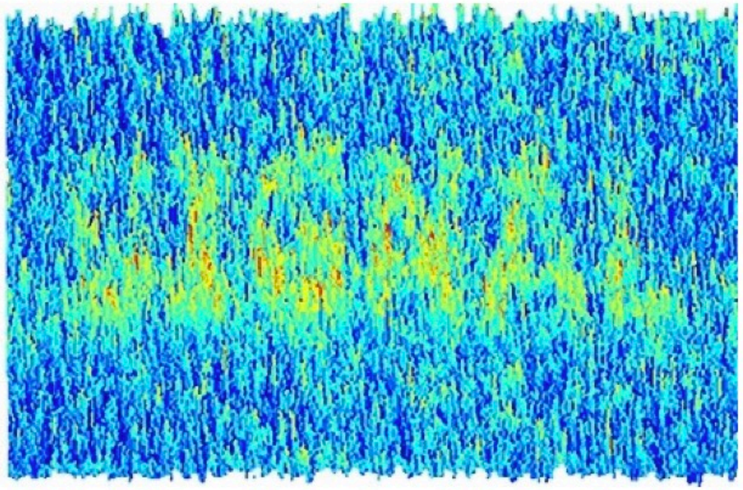
https://blog.twitter.com/engineering/en_us/a/2015/detecting-and-avoiding-bucket-imbalance-in-ab-tests.html

Source: https://hookedondata.org/guidelines-for-ab-testing/

Guidelines for A/B testing

6. Only include meaningful users in your sample

a. If you're testing a change on your homepage, then be sure the people in the experiment visit the homepage
b. Similarly, if you're lowering free shipping for people who spend a certain amount from $35 to $25, only include people with cart sizes between $25 and $35
c. Issue? Including others adds *noise* to the data, limits your ability to detect and effect



https://medium.com/@edans/twitter-and-the-importance-of-signal-to-noise-ratio-94bca0154303

Source: https://hookedondata.org/guidelines-for-ab-testing/

Guidelines for A/B testing

7. Keep the analytical approach simple

a. In most cases, simple proportion statistical tests will do the trick (i.e. Fisher's Exact, t-test, Chi-squared test)
b. If you *need* something more complex, be sure you actually need it....or maybe you just need to redesign your experiment

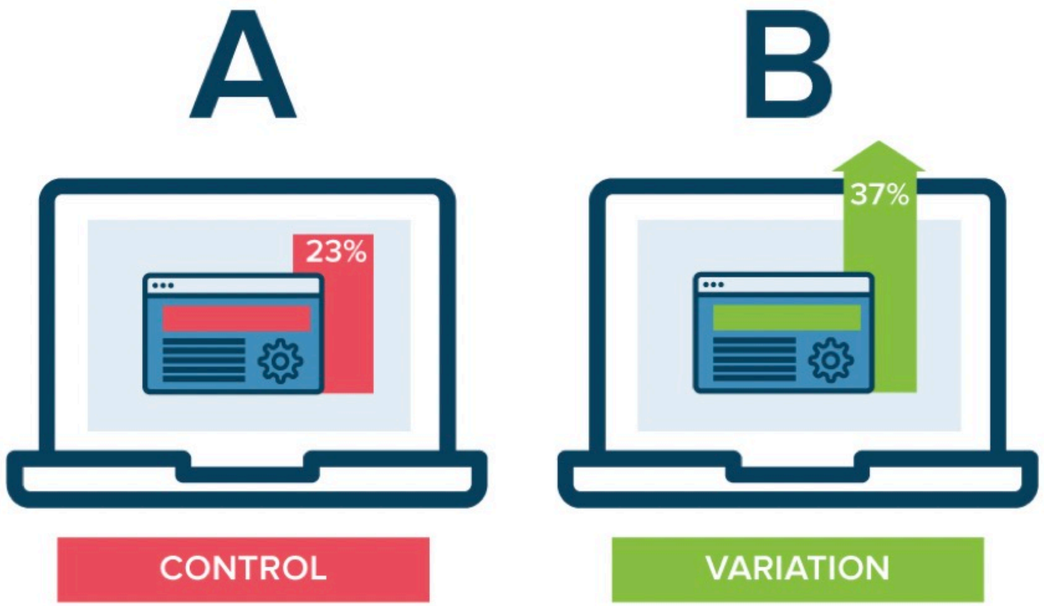
UC San Diego

Source: https://hookedondata.org/guidelines-for-ab-testing/

Guidelines for A/B testing

8. Change one thing at a time

a. Tempting to test huge differences, but best to change one small thing at a time
b. Experimental Design 101: Change one variable at a time
c. Issue? If you change more than one, you can't identify which variable caused the change



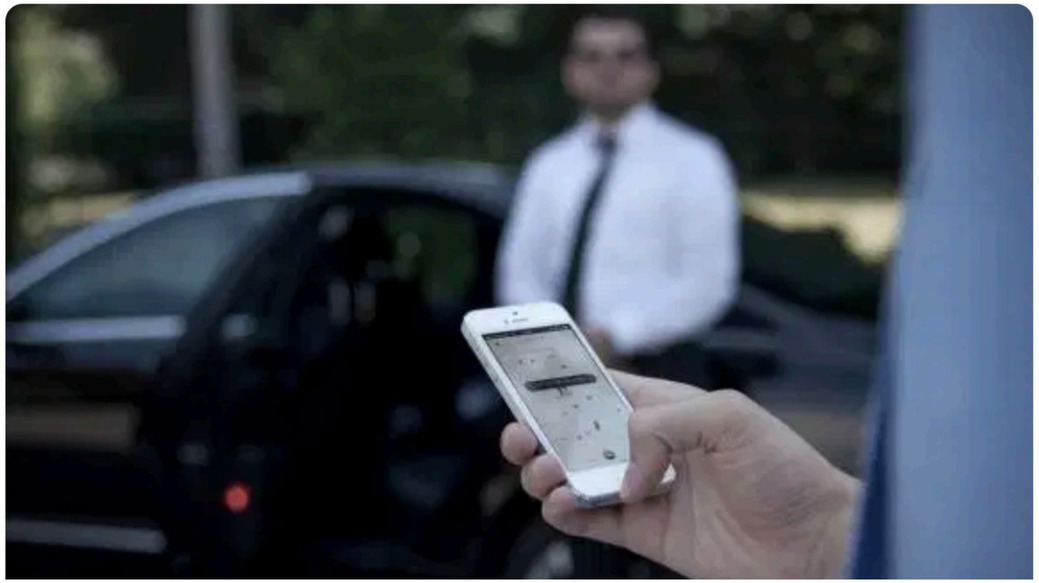
Source: https://hookedondata.org/guidelines-for-ab-testing/

A/B testing case study

Big Data

Uber simulated a city to teach drivers how to optimize their earnings

LARRY ZHOU AUGUST 11, 2014 1:17 PM



"In a blog post today, Uber data scientist Bradley Voytek explains how Uber's "science team" simulated a city and learned that taxi drivers can just stay parked between trips and make twice as much as those who drive around in search of passengers."

Source: https://venturebeat.com/2014/08/11/uber-simulated-a-city-to-teach-drivers-how-to-optimize-their-earnings/

A/B testing case study

Why?

A/B testing case study

Why are we doing this? At its core, Uber as a software platform currently has two main goals:

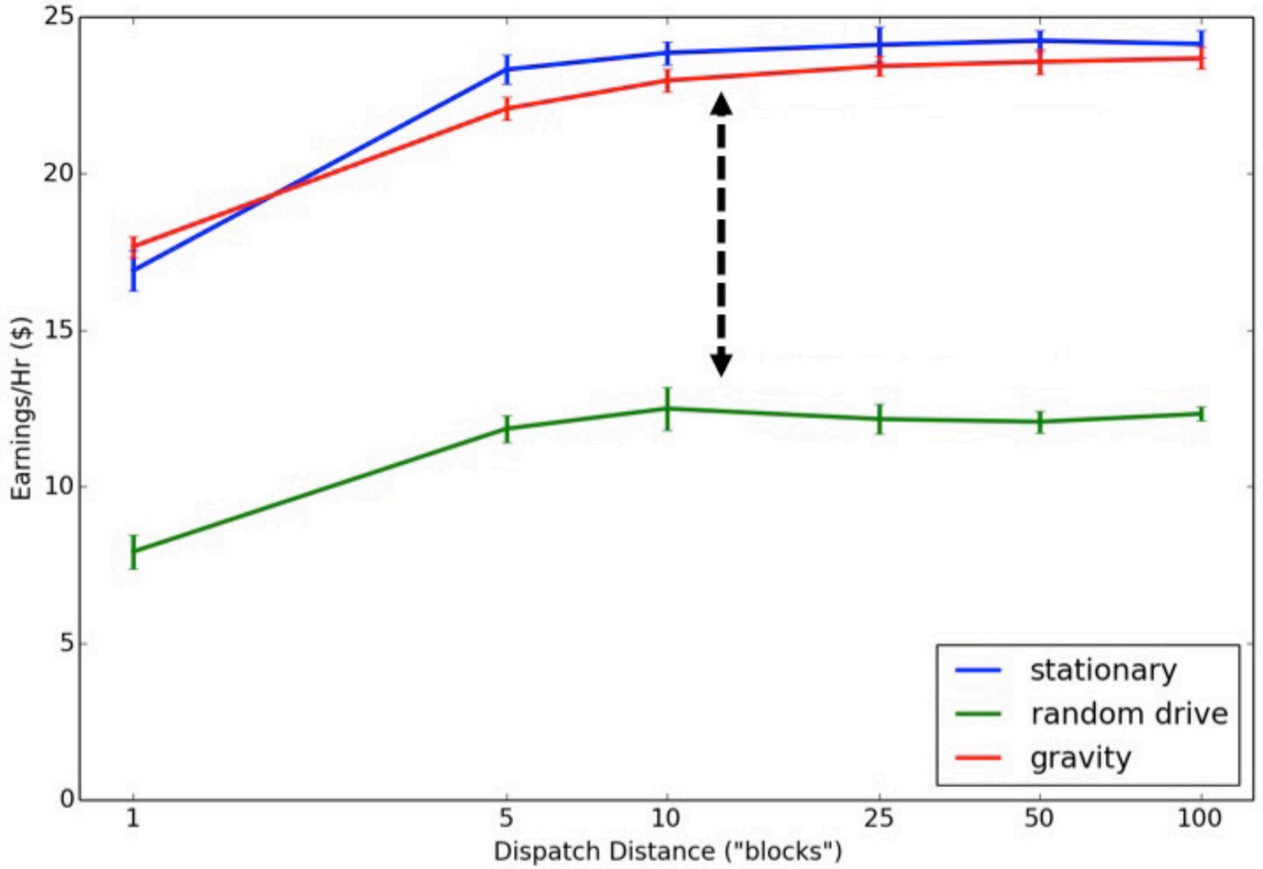
1. On the rider side, getting you a ride when you need it.
2. On the driver side, maximize trips taken on the system, which maximizes our driver partners' earnings.

> Optimizing these is *the* job for the Science Team — we want to maximize the chance of getting a ride *right* when you want one, all the while moving more people more places.

Fulfilling these goals requires modeling a number of complex, non-linear, interacting systems. This is difficult, to say the least, so what is a nerd to do? Have computers do the work for us!

Source: https://www.uber.com/newsroom/semi-automated-science-using-an-ai-simulation-framework

A/B testing case study



There are optimal dispatch distances for pairing a driver with a passenger, and there are optimal behaviors for drivers to take between trips. When dispatch distances are very short drivers should navigate back toward demand density. However when dispatch distances are relatively longer, drivers maximize their earnings by using less gas by remaining stationary between trips.

Source: https://www.uber.com/newsroom/semi-automated-science-using-an-ai-simulation-framework

COGS 9
Introduction to Data Science

EDA

UC San Diego

Today's Learning Objective

- *Explain why EDA is carried out*
- *Explain what EDA steps you would carry out given a dataset*

Data analysis

Exploratory Data Analysis

Summarize data main characteristics through the use of summaries or plots. Used to make discoveries, motivate analysis approaches or convey a message.

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

John Tukey

Exploratory:

The goal is to find unknown relationships between the variables you have measured in your data set. Exploratory analysis is open ended and designed to verify expected or find unexpected relationships between measurements.

The Data Science Process

1. Define the question you want to ask the data
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

Source: https://simplystatistics.org/2015/03/17/data-science-done-well-looks-easy-and-that-is-a-big-problem-for-data-scientists/

EDA ~ detective work

according to John Tukey (1977)

Exploratory Data Analysis (EDA) answers the question "*What can the data tell us?*"

Why EDA?

- Understand data properties
- Discover Patterns
- Generate & Frame Hypothesis
- Suggest modeling strategies
- Check assumptions (sanity checks)
- Communicate results (present the data)

Note: EDA plots do NOT have to be super-polished. They just have to be accurate.

.....and if you don't, you'll regret it

EDA

" EDA tries to go beyond descriptive statistics in the direction of true science, that is, in suggesting scientific hypotheses."

-I.J. Good (1983)

EDA

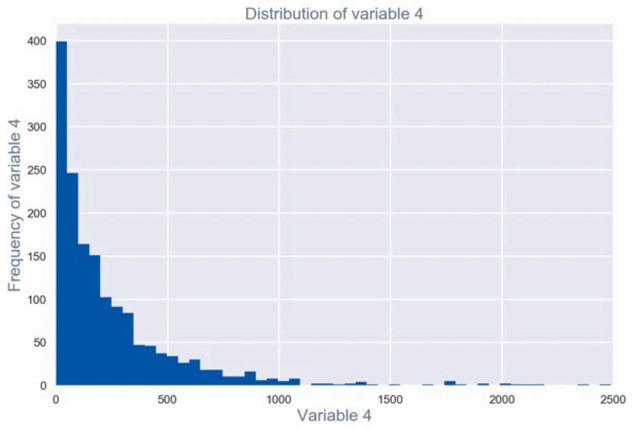
The general principles of exploratory analysis:

- Look for missing values
- Look for outlier values
- Calculate numerical summaries
- Generate plots to explore relationships
- Use tables to explore relationships
- If necessary, transform variables

EDA

EDA Approaches to "Get a Feel for the Data"

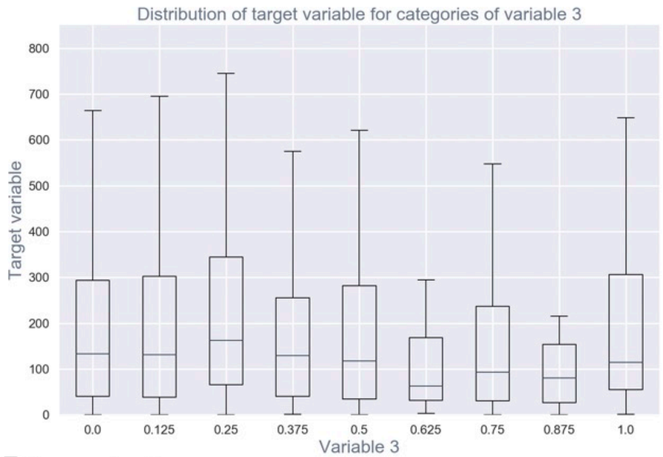
Understanding the relationship between variables in your dataset



Univariate

understanding a single variable

i.e.: histogram, densityplot, barplot



Bivariate

understanding relationship between 2 variables

i.e.: boxplot, scatterplot, grouped barplot, boxplot

Dimensionality Reduction

projecting high-D data into a lower-D space

i.e.: PCA, ICA, Clustering

Image source: https://www.svds.com/value-exploratory-data-analysis/

EDA

What is the most popular favorite color of COGS 9 students?

(univariate)

EDA

height	DOB	hair_color	zip_code	hometown	ethnicity	gender	GPA	topics	technologies	programming_languages	notes	statistics	programming	lecture_attendance
67.000000	1/26/99	black	545001	China	Asian	male	55.000	data science, con...	Canvas, Gradescope,...	Python, MATLAB, Java, R, None	Electronically (not...	6	4	I prefer to attend lecture
69.000000	11/4/95	black	*NA*	Ho Chi Minh Cit...	Asian	Male	*NA*	data science, des...	Microsoft Excel, Goo...	Python, Java	Pen and Paper (no...	7	7	I prefer to attend lecture
69.000000	11/4/95	black	*NA*	Ho Chi Minh Cit...	Asian	Male	*NA*	data science, des...	Microsoft Excel, Goo...	Python, Java	Pen and Paper (no...	7	7	I prefer to attend lecture
64.000000	12/12/00	black	0	China	Asian	Female	3.900	data science, stat...	Microsoft Excel	Java	Electronically (not...	8	2	I prefer to attend lecture
76.000000	7/25/98	black	51800	China	Asian	male	3.900	normal distributi...	Canvas, Gradescope,...	MATLAB, C, C++	Pen and Paper (no...	8	8	I prefer to attend lecture
69.000000	5/16/01	brown	91206	Glendale, Califo...	White, not Hispa...	Male	4.310	data science, dat...	Gradescope, Piazza, i...	Python	Electronically (not...	5	5	I prefer to attend lecture
64.000000	3/14/01	black	92120	San Diego	Asian	Female	4.310	continuous varia...	Google Docs, Google...	None	Electronically (not...	5	1	I prefer to attend lecture
65.000000	1/26/01	black	94134	San Francisco	Asian	Female	4.000	categorical varia...	Canvas, Google Docs...	None	Pen and Paper (no...	5	4	I prefer to attend lecture
66.000000	8/21/00	brown	95030	Encinitas	White, not Hispa...	Male	4.300	data science, dat...	Google Docs, Micros...	Python, Java	Electronically (not...	8	4	I prefer to attend lecture
69.000000	4/10/01	black	95134	San Jose	Asian	Male	4.600	categorical varia...	iclicker, Google Docs...	Python, Java	Pen and Paper (no...	6	7	I prefer to attend lecture
75.000000	1/26/01	black	94539	Fremont, CA	Asian	Male	4.710	inference, correla...	Canvas, iclicker, Goo...	Python, Java	Pen and Paper (no...	5	4	I prefer to attend lecture
65.000000	10/22/19	black	*NA*	*NA*	Asian	Female	*NA*	categorical varia...	Canvas, Gradescope,...	Python, Java	Electronically (not...	5	6	I prefer to attend lecture
69.000000	11/30/01	black	500033	Hyderabad	Asian	Male	*NA*	data science, A|B...	Canvas, Gradescope,...	Python, R	Pen/Paper and Ele...	7	4	I prefer to attend lecture
69.000000	11/2/01	black	*NA*	*NA*	American Indian...	male	4.500	data science, cat...	Canvas, Gradescope,...	Python, C++	On slides electron...	8	10	I prefer to attend lecture
71.000000	7/10/19	black	244000	China	Asian	Male	4.100	data science, dat...	Canvas, Gradescope,...	Python	Electronically (not...	7	6	I prefer not to attend lecture (i.e. catch up later, listen...
68.000000	11/4/01	black	94539	Fremont, CA	Asian	Male	3.820	data science, rep...	Canvas, Gradescope,...	Python, Java, R, Swift, Javasc...	Electronically (not...	8	7	I prefer not to attend lecture (i.e. catch up later, listen...
68.000000	8/23/01	black	518040	Shenzhen	Asian	Male	3.800	data science, repl...	Canvas, Piazza, iclick...	Python, Java, C	Pen and Paper (no...	5	10	I prefer to attend lecture
62.000000	3/5/00	black	*NA*	Shanghai	Asian	Female	4.400	continuous varia...	Canvas, Gradescope,...	None	On slides electron...	9	1	I prefer not to attend lecture (i.e. catch up later, listen...

N = 328

UC San Diego

EDA

Note: full dataset has 24 variables

height	DOB	hair_color	zip_code
hometown	ethnicity	gender	GPA
topics	technologies	programming_languages	notes
statistics	programming	lecture_attendance	year
major	work_preference	pet	vegetarian
color	siblings	animal	tv_show

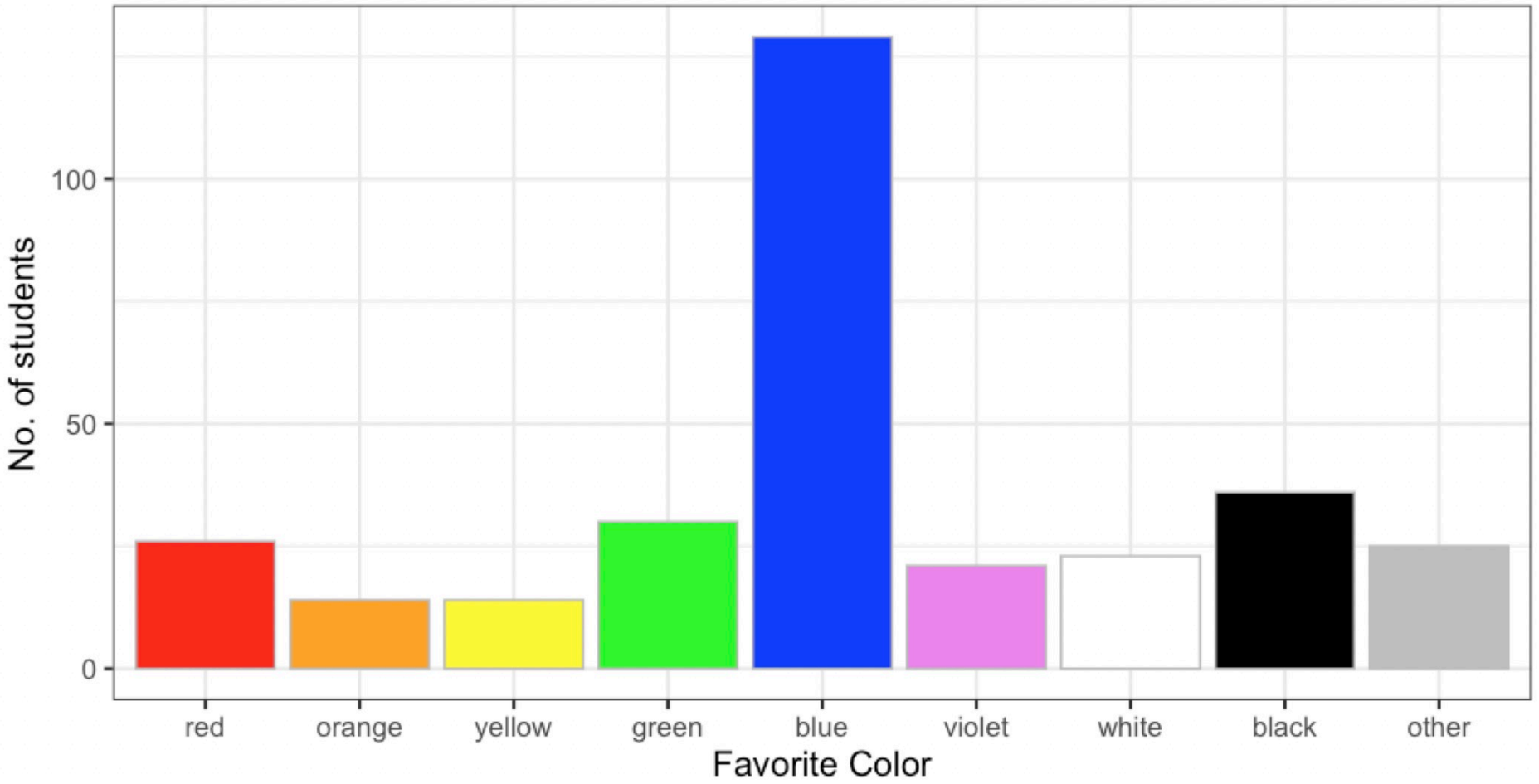
EDA

	n	percent
Aquamarine	1	0.003048780
Azure	1	0.003048780
azure / madison blue, red, violet / lavender, pastel pink, black, white	1	0.003048780
Black	36	0.109756098
Blue	129	0.393292683
Burgundy	1	0.003048780
Depends on my mood	1	0.003048780
depends on the day, hour, weather and my mood	1	0.003048780
Gray	1	0.003048780
Green	30	0.091463415
grey	1	0.003048780
I don't have a favorite color.	1	0.003048780
It switches a lot...	1	0.003048780
Lavender	2	0.006097561
Light blue	1	0.003048780
Light Pink	1	0.003048780
maroon	1	0.003048780
Orange	14	0.042682927
Pink	2	0.006097561
purple	1	0.003048780
Purple	2	0.006097561
Red	26	0.079268293
sunset orange	1	0.003048780
They’re all so fun!	1	0.003048780
too old to care about a favorite colour	1	0.003048780
turquoise	2	0.006097561
Turquoise	2	0.006097561
ultraviolet	1	0.003048780
Violet	18	0.054878049
White	23	0.070121951
Yellow	14	0.042682927
<NA>	10	0.030487805

Requires some wrangling...

UC San Diego

EDA


No. of students
100
50
0
red
orange
yellow
green
blue
violet
white
black
other
Favorite Color

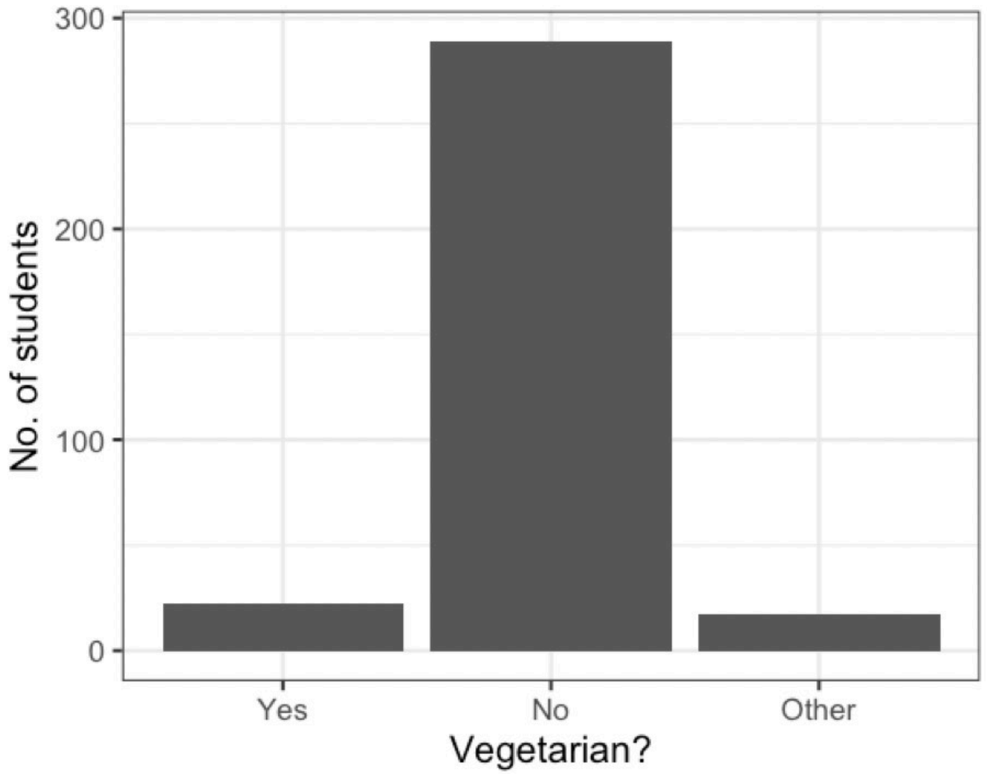
EDA

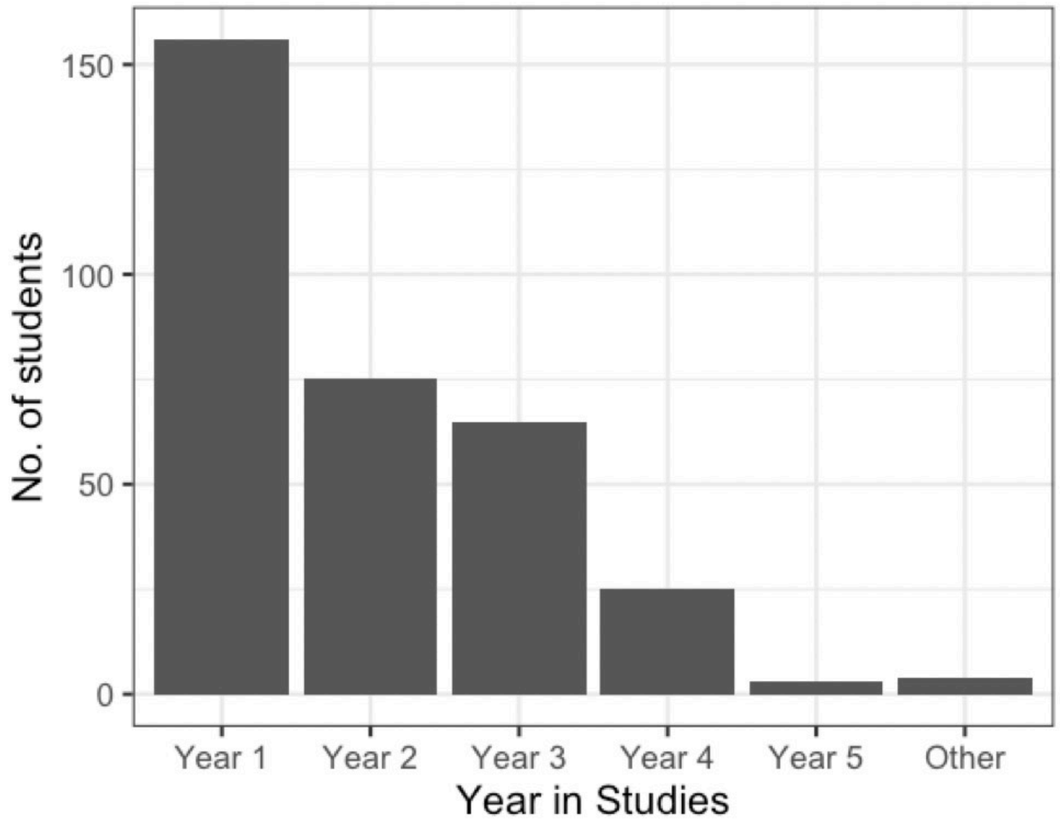
What % of COGS 9 students are vegetarian?
(univariate)

OR

What year in their studies are COGS 9 students?
(univariate)

EDA


300
200
100
0
No. of students
Yes
No
Other
Vegetarian?

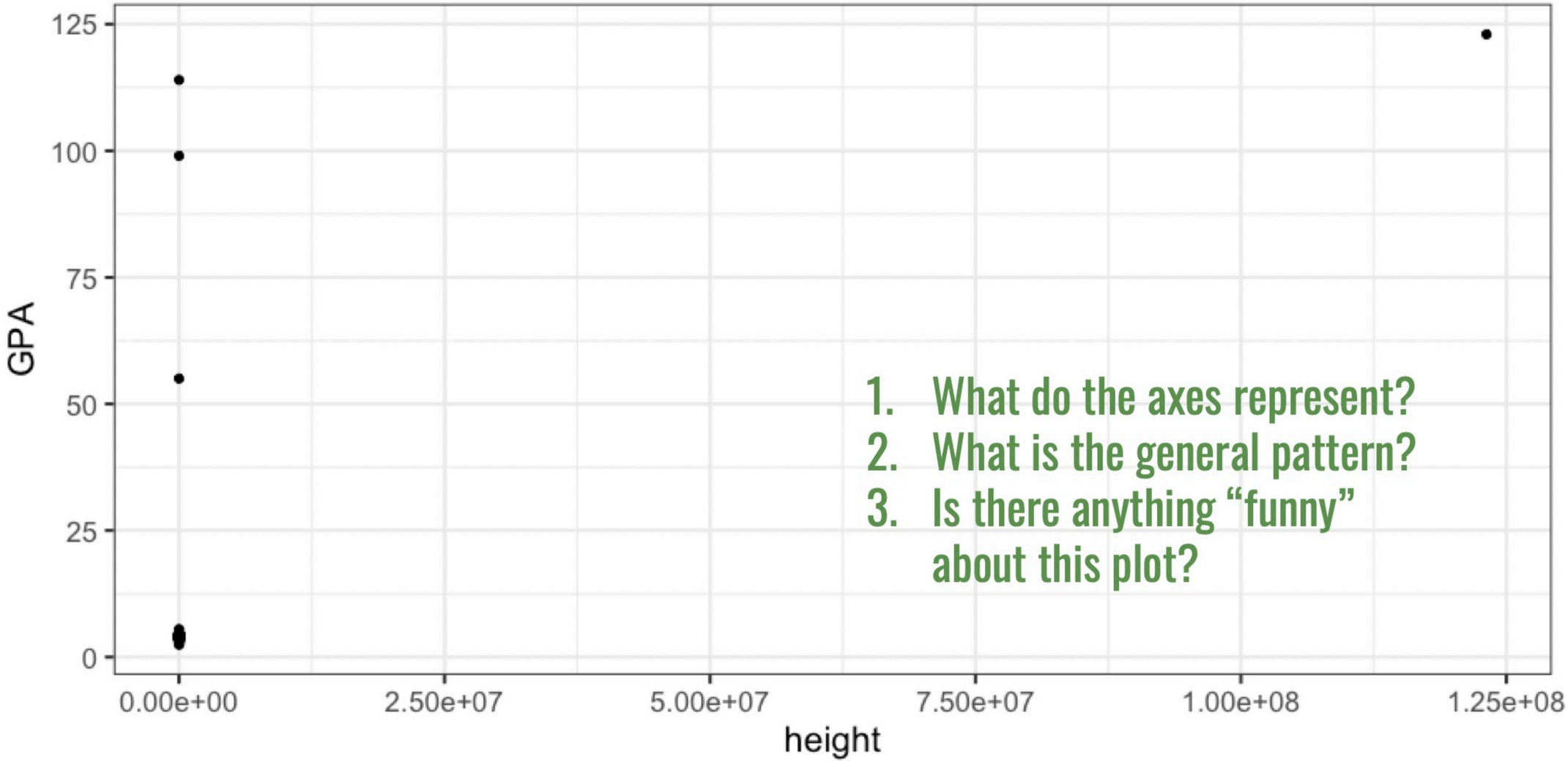

150
100
50
0
No. of students
Year 1
Year 2
Year 3
Year 4
Year 5
Other
Year in Studies

EDA

Do taller students tend to have higher GPAs?

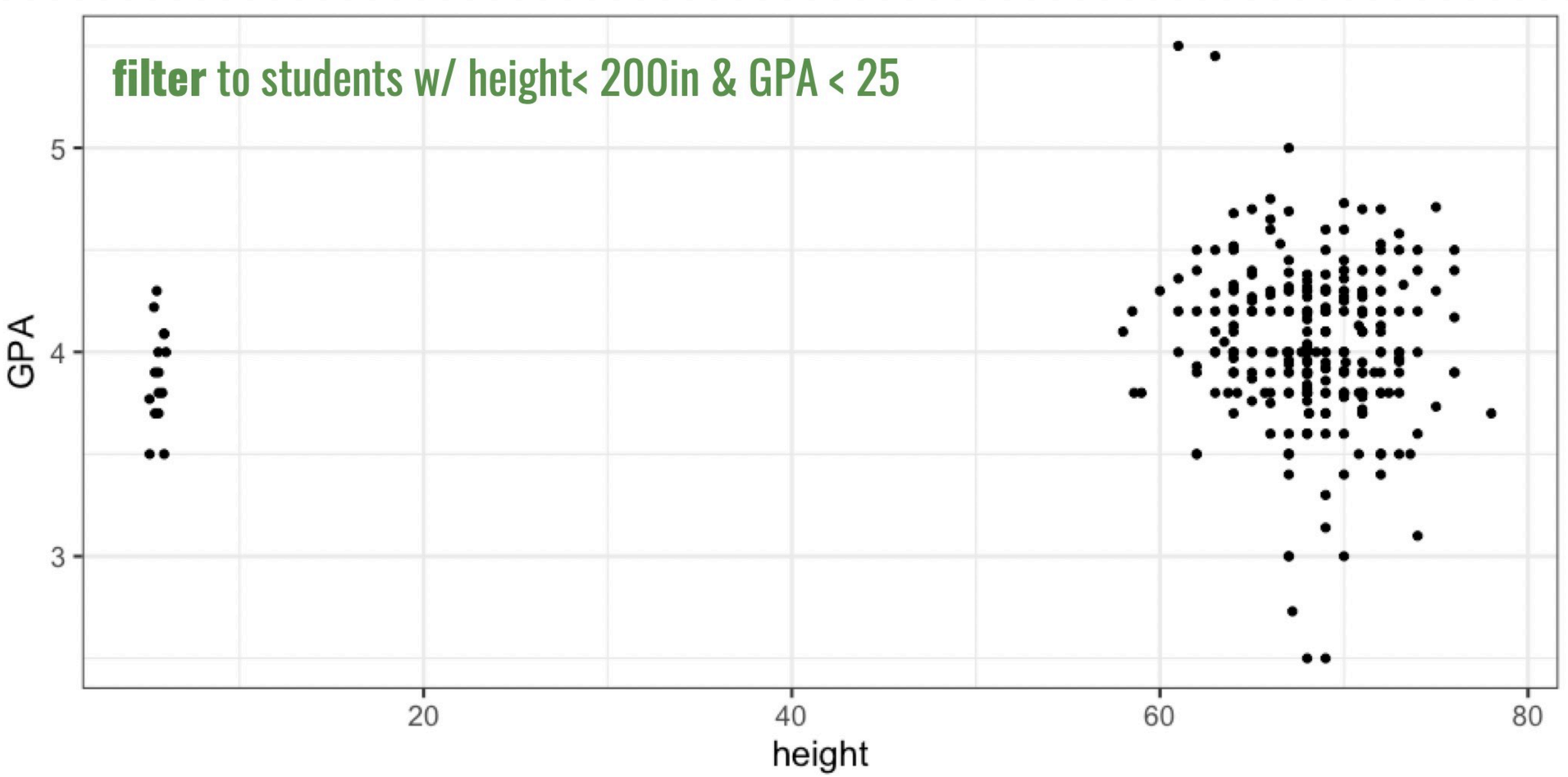
(bivariate)

EDA



1. What do the axes represent?
2. What is the general pattern?
3. Is there anything "funny" about this plot?

EDA

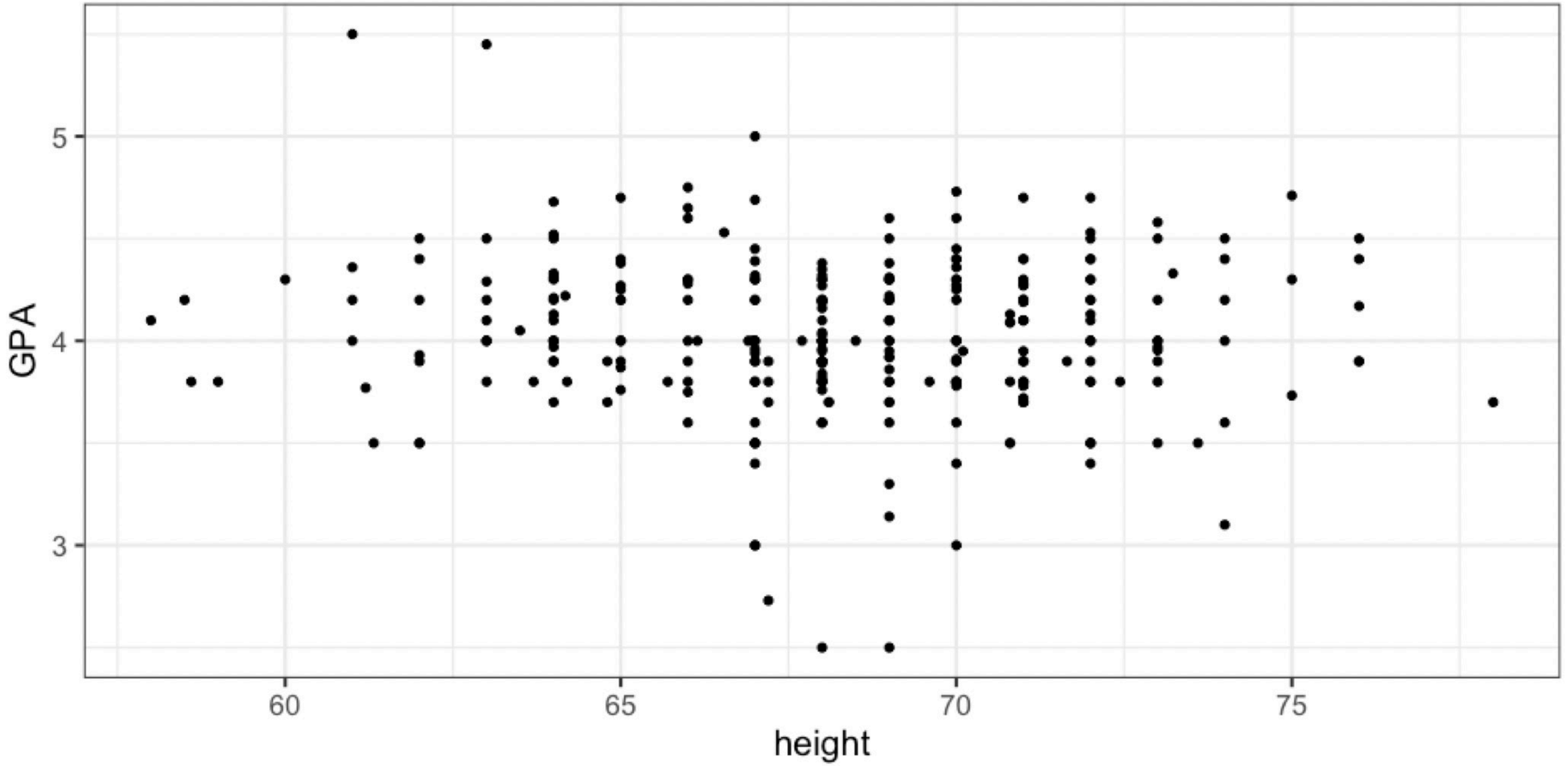

filter to students w/ height< 200in & GPA < 25
GPA
5
4
3
20
40
60
80
height

EDA

filter to students height < 20in & GPA < 25
select height and GPA

	height	GPA
	<dbl>	<dbl>
1	5.6	3.7
2	5.35	4.22
3	5.58	4
4	5.4	3.9
5	5.9	4.09
6	5.9	3.5
7	6	4
8	5.1	3.77
9	5.6	3.9
10	5.11	3.5
11	5.6	3.8
12	5.8	3.8
13	5.9	4.09
14	5.5	4.3
15	5.4	3.7

EDA



EDA

Where did COGS 9 students grow up?

(univariate, geospatial)

EDA

In what zip code (postal code) were you born?

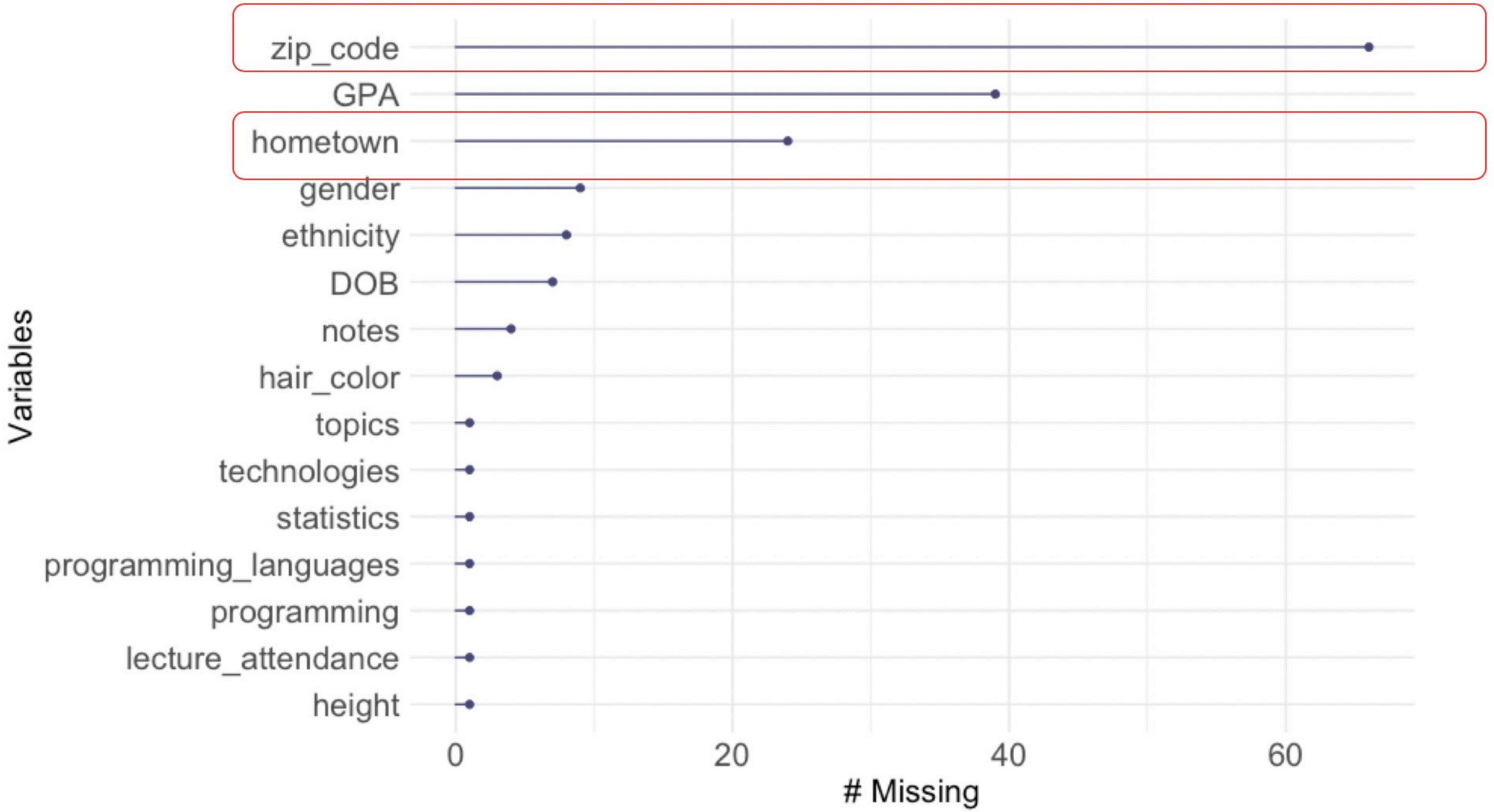
Short answer text

What is your Hometown?

Short answer text

...which should we use?

EDA


zip_code
GPA
hometown
gender
ethnicity
DOB
notes
hair_color
topics
technologies
statistics
programming_languages
programming
lecture_attendance
height
Variables
0
20
40
60
Missing

EDA

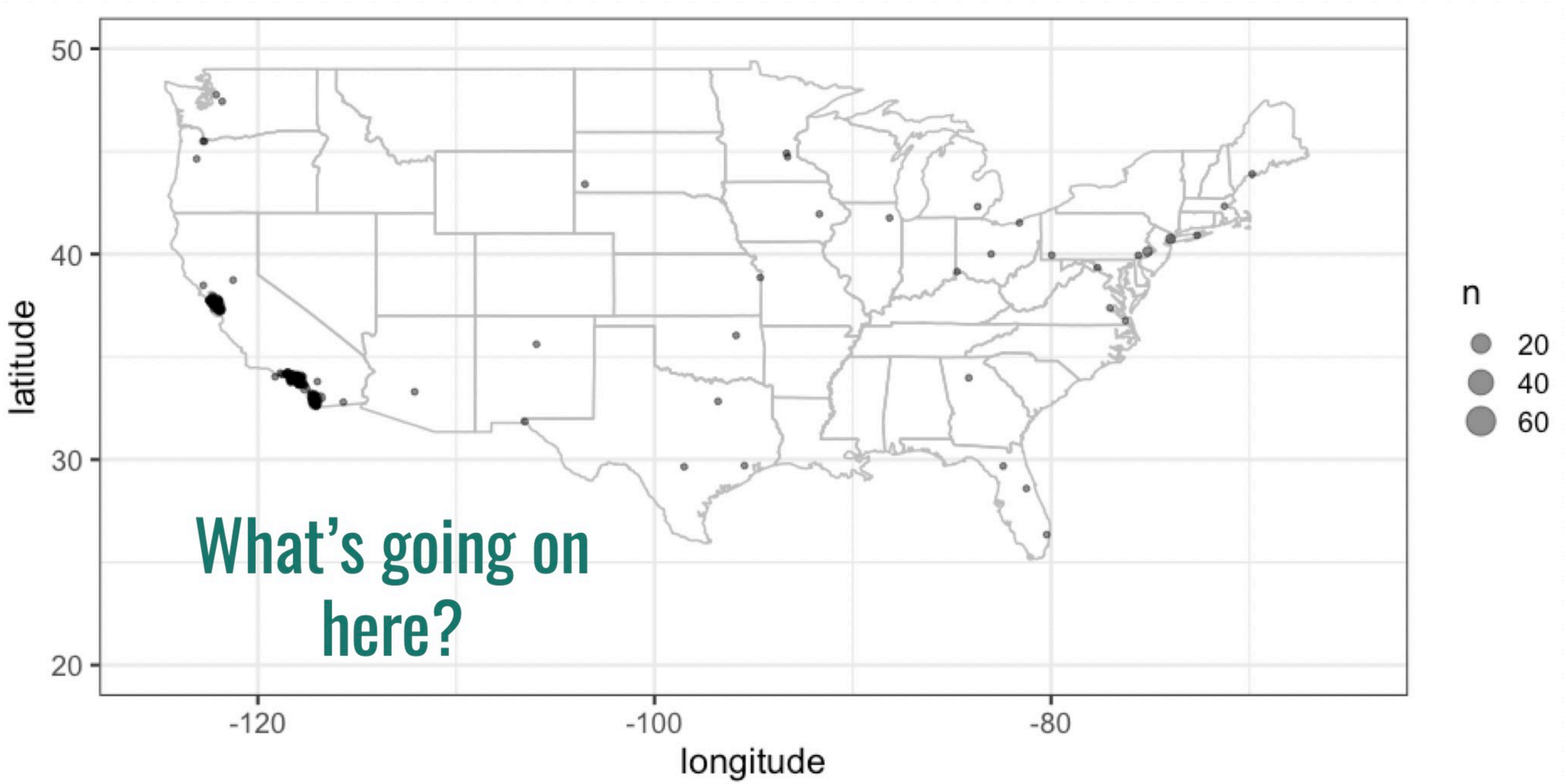
zip_code	hometown
545001	China
NA	Ho Chi Minh City, Vietnam
NA	Ho Chi Minh City, Vietnam
0	China
51800	China
91206	Glendale, California
92120	San Diego
94134	San Francisco
95030	Encinitas
95134	San Jose
94539	Fremont, CA

EDA

General approach to using zip codes:

1. Map zip codes to longitude and latitude
2. Count how many people fall into each zip code
3. Plot each place on map

EDA

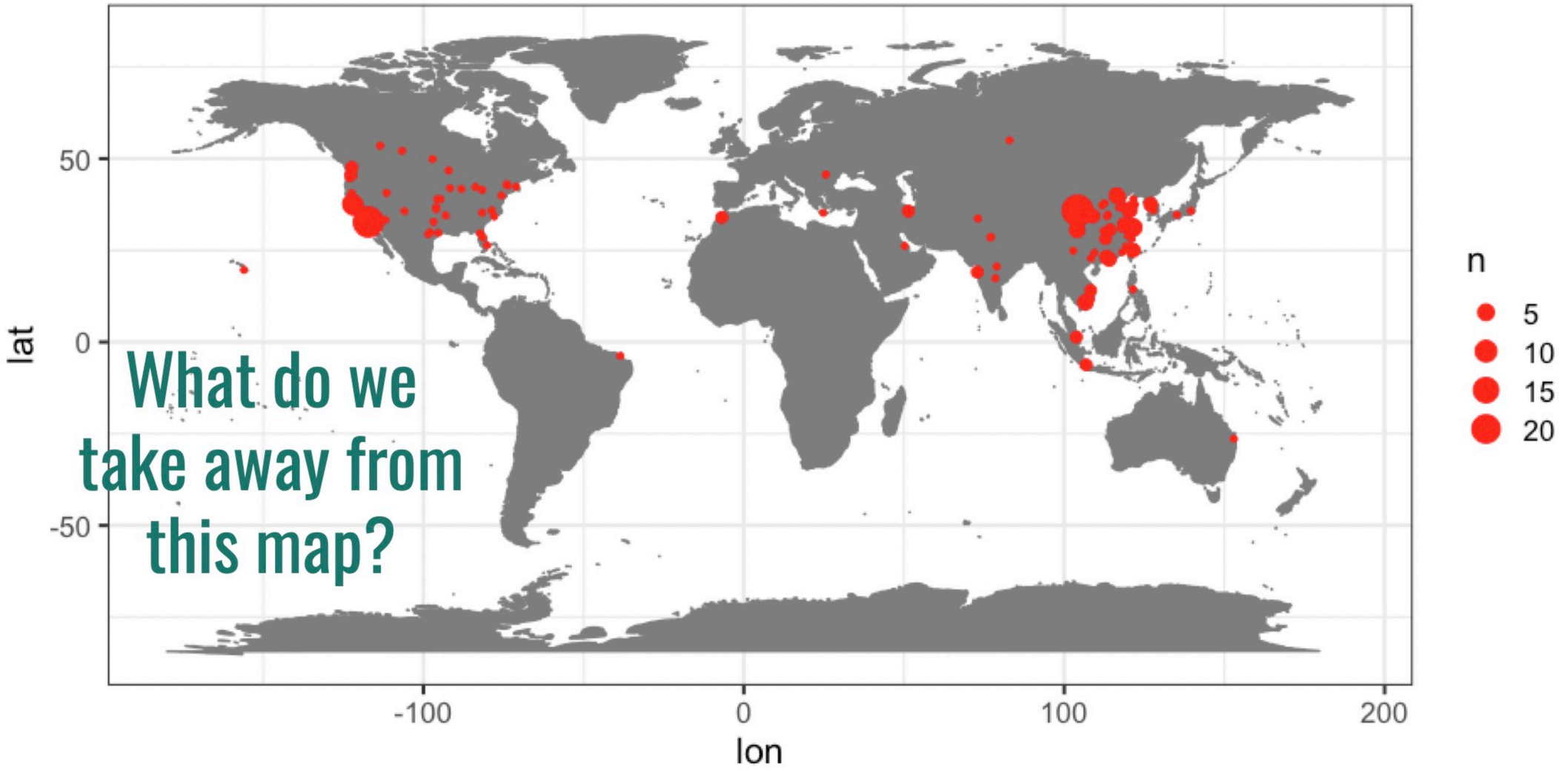

50
40
30
20
latitude
-120
-100
-80
longitude
n
20
40
60
What's going on here?

EDA

zip_code	hometown
545001	China
NA	Ho Chi Minh City, Vietnam
NA	Ho Chi Minh City, Vietnam
0	China
51800	China
91206	Glendale, California
92120	San Diego
94134	San Francisco
95030	Encinitas
95134	San Jose
94539	Fremont, CA

What if instead we used *geolocation* based on what people specified in 'hometown'?

EDA


50
0
-50
lat
-100
0
100
200
lon
n
5
10
15
20
What do we take away from this map?

EDA

What we've learned from our exploratory analysis

Things we learned:

- First years most common in COGS 9
- Most students not vegetarian
- Blue is the most popular favorite color
- Little to no relationship between GPA and height
- Zip code not the whole story
- COGS 9 students come from all over the world!

When not to do EDA

- To identify samples that you can remove from your study *after you've already analyzed the data*
- After running a statistical test and seeing that your p-value is 0.054
- After completing an analysis and getting an answer you don't like
- To *improve* the correlation between two variables

> *"If you torture the data long enough, it will confess"*
>
> - Ronald Coase (Economist)

EDA is <u>NOT</u> a tool to get your data analysis to give you the results you *want*.