

Course Reminders

Due Sunday (11:59 PM)

- D4
- Q5
- Project Proposal
- [Mid-course survey](#) (*optional for EC, link also on Canvas assignment*)
- [Weekly Project Survey](#) (*optional, link also on Canvas assignment and homepage*)

Notes:

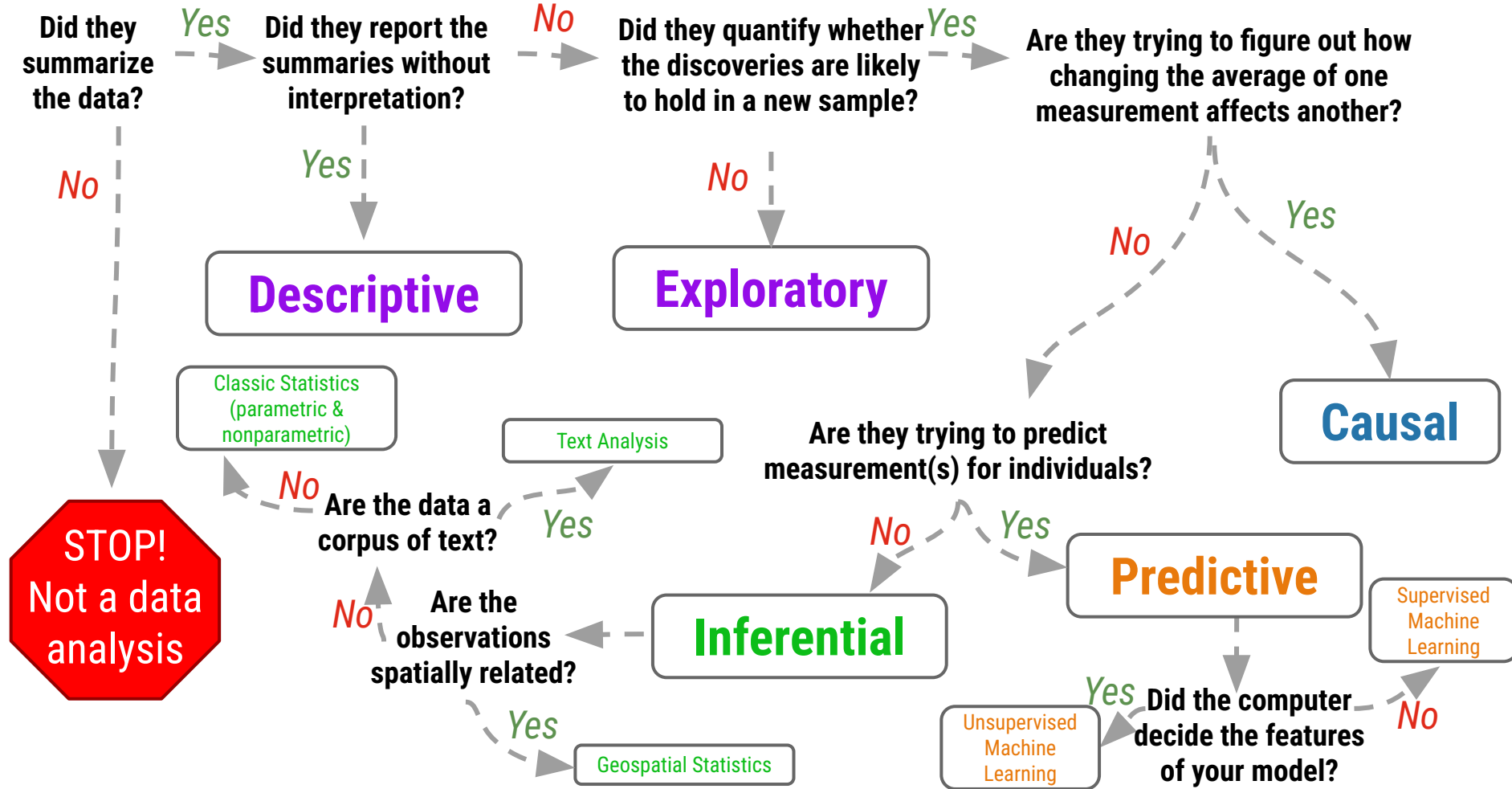
- Project grader now assigned - specified as an Issue on your project repo
- D3 scores posted

Exploratory Data Analysis

Shannon E. Ellis, Ph.D
UC San Diego



Department of Cognitive Science
sellis@ucsd.edu



Exploratory Data Analysis (EDA): The goal is to **understand your data** (understand the components of a data set, describe what they are, find unknown relationships between the variables you have measured in your data set, etc.). Exploratory analysis is **open-ended** and designed to verify expected or find unexpected relationships between measurements.

Why EDA?

- Understand data properties
 - Structure (file format)
 - Granularity (what is each observation?)
 - Scope (how complete are our data? Will they work for our needs?)
 - Temporality (situated in time?)
 - Faithfulness (how trustworthy? Gut checks - check assumptions)
- Discover Patterns
- Generate & Frame Hypothesis
- Suggest modeling strategies
- “See the data” results (start to visualize the data)

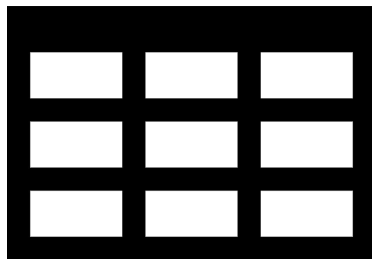
.....and if you don't, you'll regret it

It's *always* worth spending time at the beginning of a project to determine whether or not the data you have are garbage. Be certain they are actually able to help you answer the question you're interested in.

GIGO : Garbage In. Garbage Out.



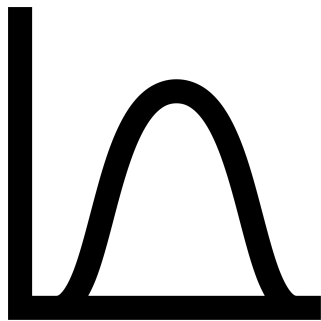
First
steps...



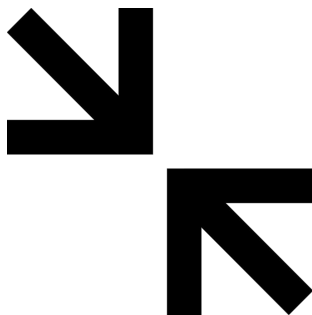
Size



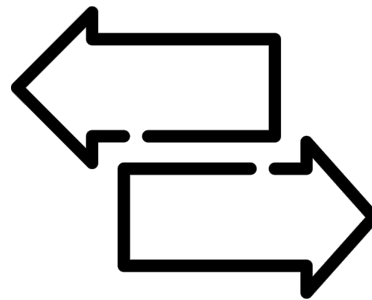
Missingness



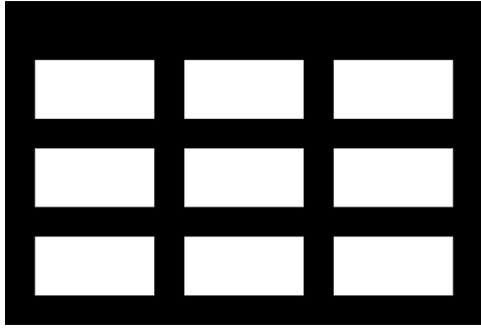
Shape



Central Tendency



Variability



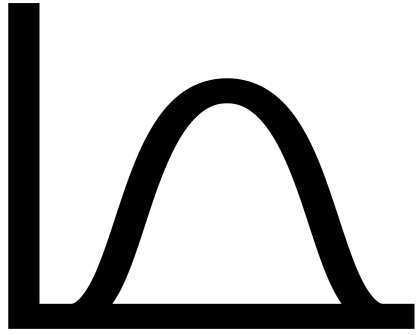
Size

How many observations (rows) and variables (columns) you have is an important first step. You should always be aware of the **size** of your dataset .



Missingness

It's critical to know **how many observations have missing data** for variables of interest in your data. Knowing *why* their missing is also important.



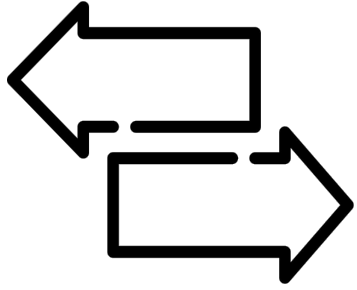
Shape

It's critical to know the distribution of the variables in your dataset. Certain statistical approaches can only be used with certain distributions.



Central Tendency

Knowing the mean, median, and/or mode can help you get an idea of what a typical value is for your variable(s) of interest



Variability

The central tendency tells you part of the story. The **variability in the values** in your observation helps fill in the rest.

Descriptive Statistics & Summary

*“We must suppress some of the truth to communicate the truth...
In short, the techniques of descriptive statistics are designed to
match the salient features of the data set to human cognitive
abilities.”*

-I.J. Good (1983)



Which of the following is NOT something accomplished by EDA?

- ☐ **A** Describes typical values in your dataset
- ☐ **B** Determines the size of your dataset
- ☐ **C** Establishes causal relationships between variables
- ☐ **D** Identifies missing data
- ☐ **E** Determines how variable values in your dataset are

In academic publishing and data science reports “Table 1” often describes the data

Characteristic	Ranibizumab Monthly (N = 301)	Bevacizumab Monthly (N = 286)	Ranibizumab as Needed (N = 298)	Bevacizumab as Needed (N = 300)
Age — no. (%)				
50–59 yr	2 (0.7)	1 (0.3)	6 (2.0)	2 (0.7)
60–69 yr	33 (11.0)	28 (9.8)	31 (10.4)	34 (11.3)
70–79 yr	102 (33.9)	84 (29.4)	115 (38.6)	103 (34.3)
80–89 yr	142 (47.2)	150 (52.4)	126 (42.3)	142 (47.3)
≥90 yr	22 (7.3)	23 (8.0)	20 (6.7)	19 (6.3)
Mean — yr	79.2±7.4	80.1±7.3	78.4±7.8	79.3±7.6
Sex — no. (%)				
Female	183 (60.8)	180 (62.9)	185 (62.1)	184 (61.3)
Male	118 (39.2)	106 (37.1)	113 (37.9)	116 (38.7)
Race — no. (%)†				
White	297 (98.7)	281 (98.3)	296 (99.3)	294 (98.0)
Other	4 (1.3)	5 (1.7)	2 (0.7)	6 (2.0)
History of myocardial infarction — no. (%)	34 (11.3)	40 (14.0)	30 (10.1)	36 (12.0)
History of stroke — no. (%)	14 (4.7)	18 (6.3)	22 (7.4)	16 (5.3)
History of transient ischemic attack — no. (%)	12 (4.0)	25 (8.7)	12 (4.0)	19 (6.3)
Blood pressure — mm Hg				
Systolic	134±18	135±19	136±17	135±17
Diastolic	75±10	75±10	76±9	75±10
Visual-acuity score and Snellen equivalent				
68–82 letters, 20/25–40 — no. (%)	111 (36.9)	94 (32.9)	116 (38.9)	103 (34.3)
53–67 letters, 20/50–80 — no. (%)	98 (32.6)	118 (41.3)	108 (36.2)	119 (39.7)
38–52 letters, 20/100–160 — no. (%)	67 (22.3)	53 (18.5)	58 (19.5)	58 (19.3)
23–37 letters, 20/200–320 — no. (%)	25 (8.3)	21 (7.3)	16 (5.4)	20 (6.7)
Mean score	60.1±14.3	60.2±13.1	61.5±13.2	60.4±13.4
Total thickness at fovea — μm‡	458±184	463±196	458±193	461±175
Retinal thickness plus subfoveal-fluid thickness at fovea — μm	251±122	254±121	247±122	252±115
Foveal center involvement — no. (%)				
Choroidal neovascularization	176 (58.5)	153 (53.5)	176 (59.1)	183 (61.0)
Fluid	85 (28.2)	81 (28.3)	77 (25.8)	72 (24.0)
Hemorrhage	20 (6.6)	24 (8.4)	24 (8.1)	25 (8.3)
Other	18 (6.0)	20 (7.0)	15 (5.0)	18 (6.0)
No choroidal neovascularization or not possible to grade	2 (0.7)	8 (2.8)	6 (2.0)	2 (0.7)

^a Plus-minus values are means ±SD.
[†] Race was self-reported.
[‡] Total thickness at the fovea includes the retina, subretinal fluid, choroidal neovascularization, and retinal pigment epithelial elevation.

Table 1. Baseline Characteristics of the Patients.*

Characteristic	Ranibizumab Monthly (N = 301)	Bevacizumab Monthly (N = 286)	Ranibizumab as Needed (N = 298)	Bevacizumab as Needed (N = 300)
Age — no. (%)				
50–59 yr	2 (0.7)	1 (0.3)	6 (2.0)	2 (0.7)
60–69 yr	33 (11.0)	28 (9.8)	31 (10.4)	34 (11.3)
70–79 yr	102 (33.9)	84 (29.4)	115 (38.6)	103 (34.3)
80–89 yr	142 (47.2)	150 (52.4)	126 (42.3)	142 (47.3)
≥90 yr	22 (7.3)	23 (8.0)	20 (6.7)	19 (6.3)
Mean — yr	79.2±7.4	80.1±7.3	78.4±7.8	79.3±7.6
Sex — no. (%)				
Female	183 (60.8)	180 (62.9)	185 (62.1)	184 (61.3)
Male	118 (39.2)	106 (37.1)	113 (37.9)	116 (38.7)
Race — no. (%)†				
White	297 (98.7)	281 (98.3)	296 (99.3)	294 (98.0)
Other	4 (1.3)	5 (1.7)	2 (0.7)	6 (2.0)

* Plus-minus values are means ±SD.

† Race was self-reported.

‡ Total thickness at the fovea includes the retina, subretinal fluid, choroidal neovascularization, and retinal pigment epithelial elevation.

Size

Shape
Central
tendency
variability

Zooming in on this we
see variables
stratified by Age, Sex,
and Race