

Course Reminders

Due Dates:

- Project Proposal due Sunday
- A2 due *next* Sunday

Notes:

- Group update emails were sent out
- If you've not gotten a reply from a team member by Wednesday, please email me.
- Course survey open until Wednesday

Approaches to Analysis

Shannon E. Ellis, Ph.D
UC San Diego

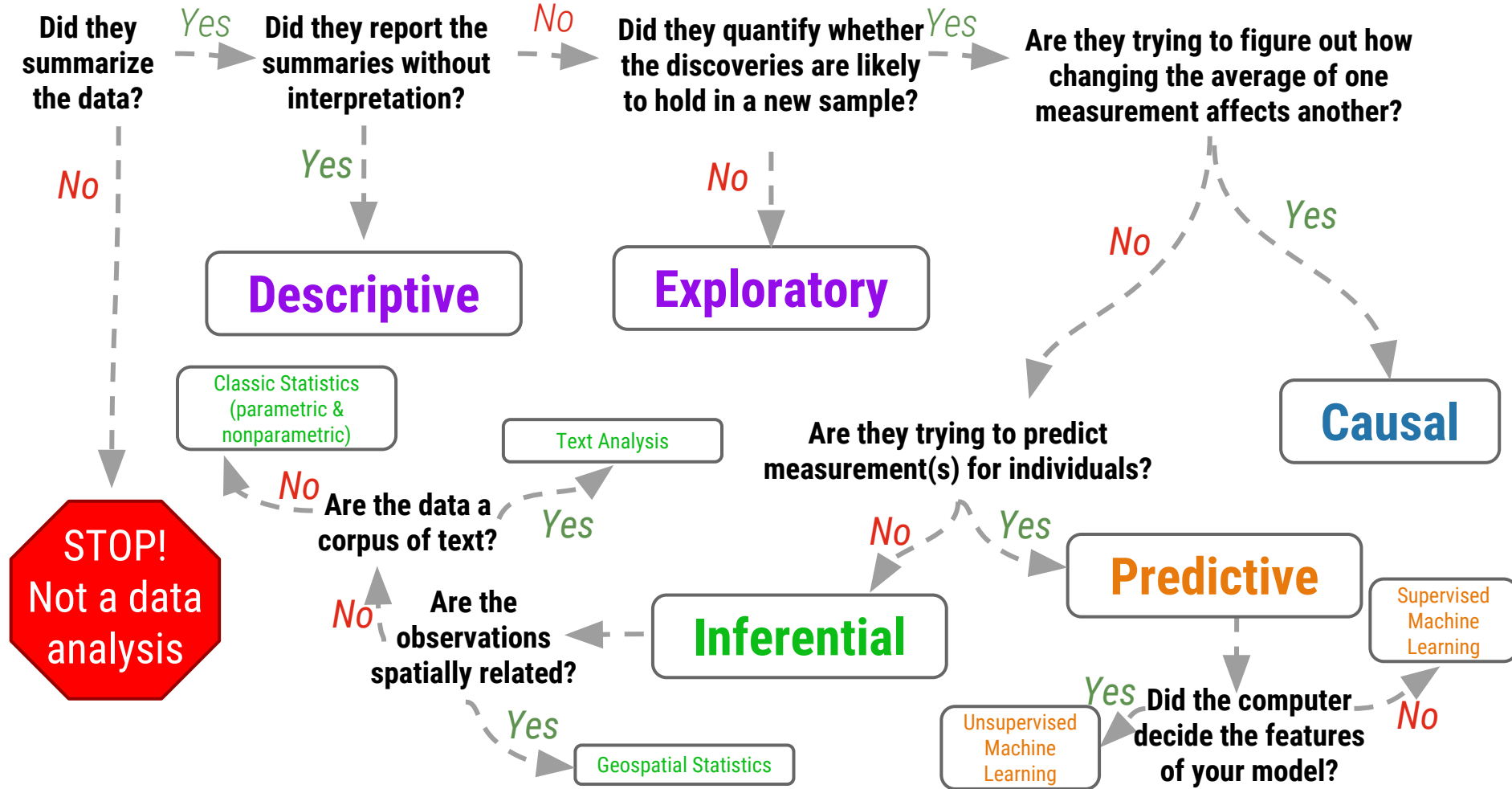


Department of Cognitive Science
sellis@ucsd.edu

*“Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, **analyzing the data**, and communicating the answer to the question to a relevant audience.”*



To do this, you have to look at, describe, and explore the data



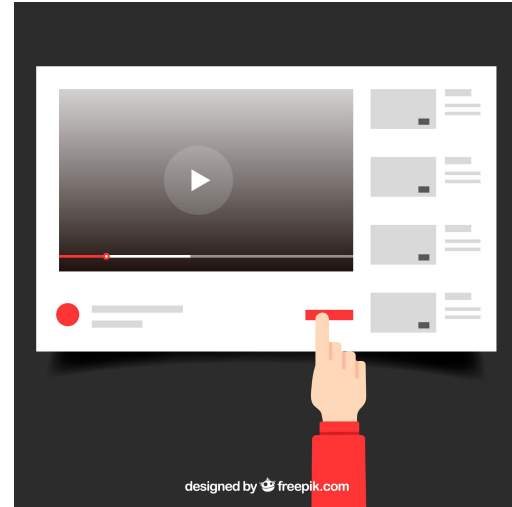
Summary: Analytical Approaches

1. **Descriptive** (and **Exploratory**) Data Analysis are the first step(s)
2. **Inference** establishes relationships
 - a. Classic Statistics
 - b. Geospatial Analysis
 - c. Text Analysis
3. Machine Learning is for **prediction**
 - a. Supervised
 - b. Unsupervised
4. Experiments best way to establish **causality**

Descriptive: The goal of descriptive analysis is to understand the components of a data set, describe what they are, and explain that description to others who might want to understand the data.

- **Problem:** Understanding whether users are nice or mean on Youtube
- **Data science question:** Are the words that people use in their comments more frequently positive words (great, awesome, nice, useful) or negative words (bad, stupid, lame, awful)?
- **Type of analysis:** Descriptive analysis

To answer this you
would calculate
statistics about YouTube
comments



Statistics

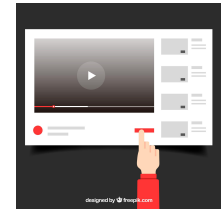
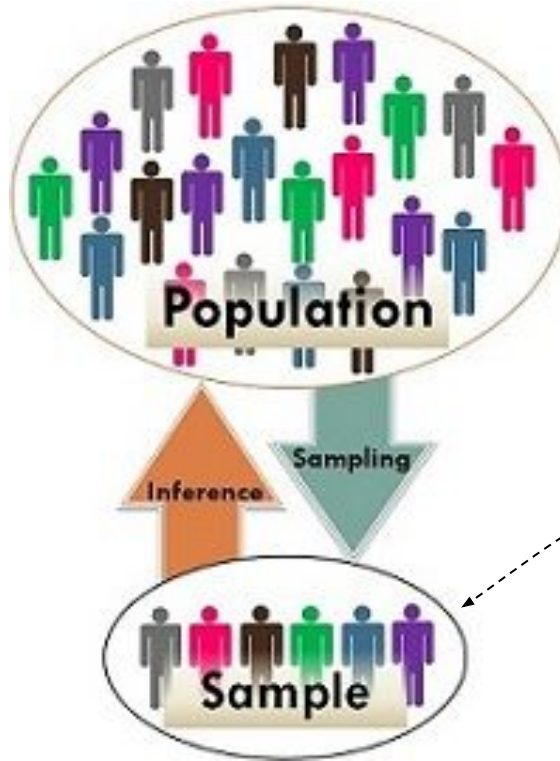
*“the science that deals with the **collection, classification, analysis, and interpretation of numerical facts or data**”*

statistic

“A quantity computed from a sample”

Populations & Samples

We want to learn something about this...



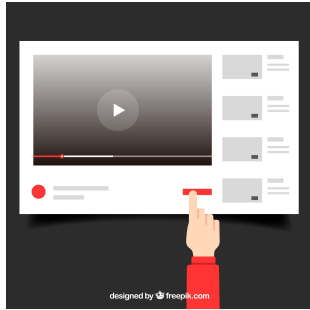
Our population: all YouTube comments

Our sample: 100,000 comments

...but we can only *actually* collect data from this

statistic

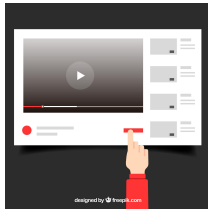
“A quantity computed from a sample”



For our YouTube analysis, we could take a random sample of comments from YouTube and calculate the following statistic: *the number of positive and the number of negative words in each review.*

Best sampling practices:

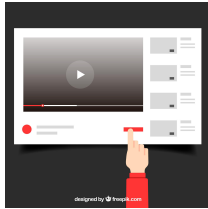
- Always think about what your population is
- Collect data from a sample that is representative of your population
- If you have no choice but to work with a dataset that is not collected randomly and is biased, be careful not to generalize your results to the entire population



You'd want to be sure you sample randomly across *all* YouTube comments, making sure not to get more comments from one genre over another, or one location over another, etc.

Examples of bad sampling:

- Surveying subscribers of a gun-related magazine for research on Americans' attitudes toward owning guns
- Randomly sampling Facebook users for what TV shows people like



To understand *all* YouTube comments, you wouldn't just want to sample from one YouTube channel, or videos in a single language.

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.

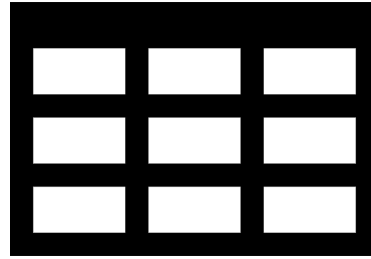




For the survey data I collected from you all, which of the following best describes the population I could generalize findings back to.

- ☐ **A** Undergraduates
- ☐ **B** Undergraduates in the US
- ☐ **C** Undergraduates at UCSD
- ☐ **D** Students aged 18-25
- ☐ **E** UCSD COGS108 students

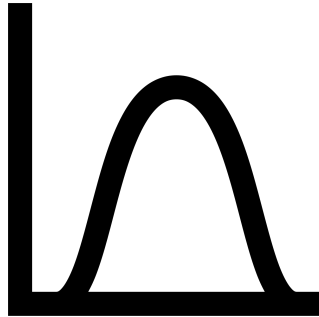
Descriptive Analysis



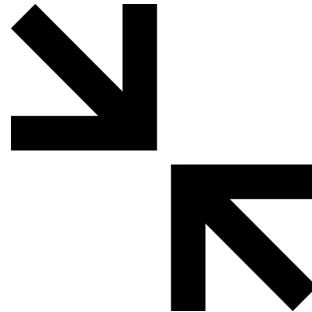
Size



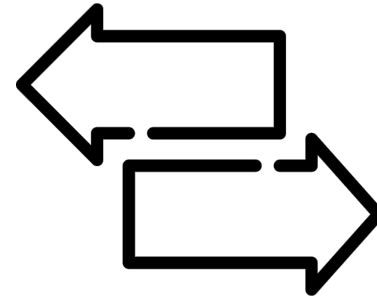
Missingness



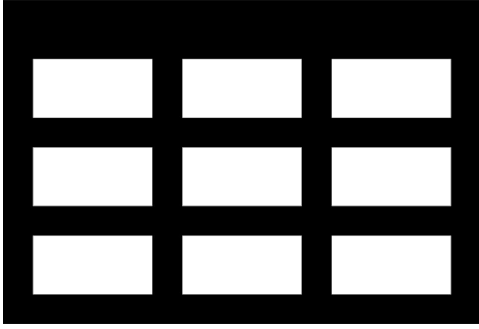
Shape



Central Tendency



Variability



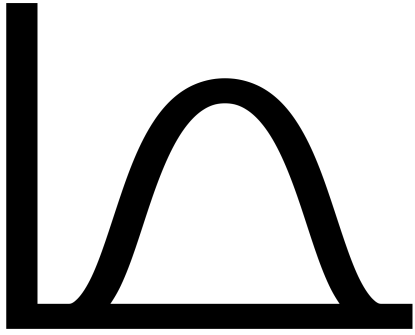
Size

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.



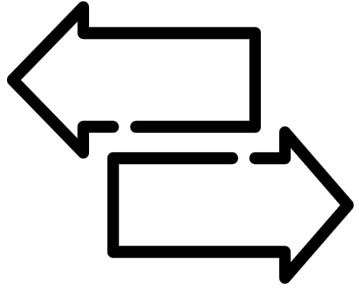
Shape

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.



Which of the following is NOT something accomplished by a descriptive analysis?

- ☐ **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

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)

Descriptive Analyses
are often included as
“Table 1” in academic
publications

Table 1. Baseline Characteristics of the Patients.^a

Characteristic	Ranibizumab Monthly (N = 301)	Bevacizumab Monthly (N = 286)	Ranibizumab as Needed (N = 298)	Becavizumab 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)

* 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

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.

Shape
Central
tendency

variability

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

Descriptive Statistics & Summary

Calculating descriptive statistics, understanding what they tell you about your data, and reporting them are critical steps in every analysis.

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.

Exploratory



Exploratory Data Analysis (EDA)
detective work answering 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)

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

Exploratory

YOU MUST ALWAYS EXPLORE YOUR DATA

衆瞽
探象之圖



The
dataset

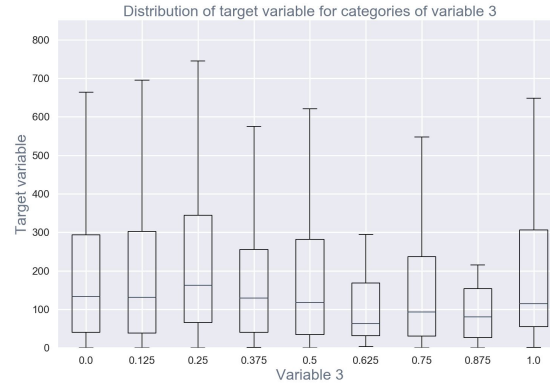
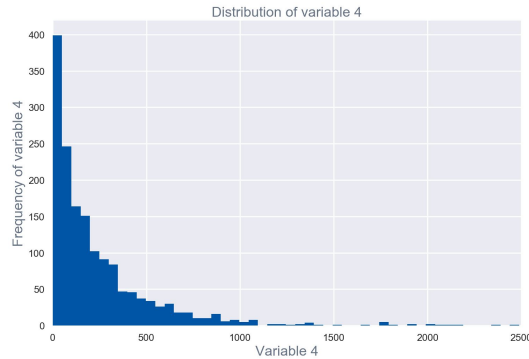
You

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 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



How does EDA differ from Descriptive Analysis?

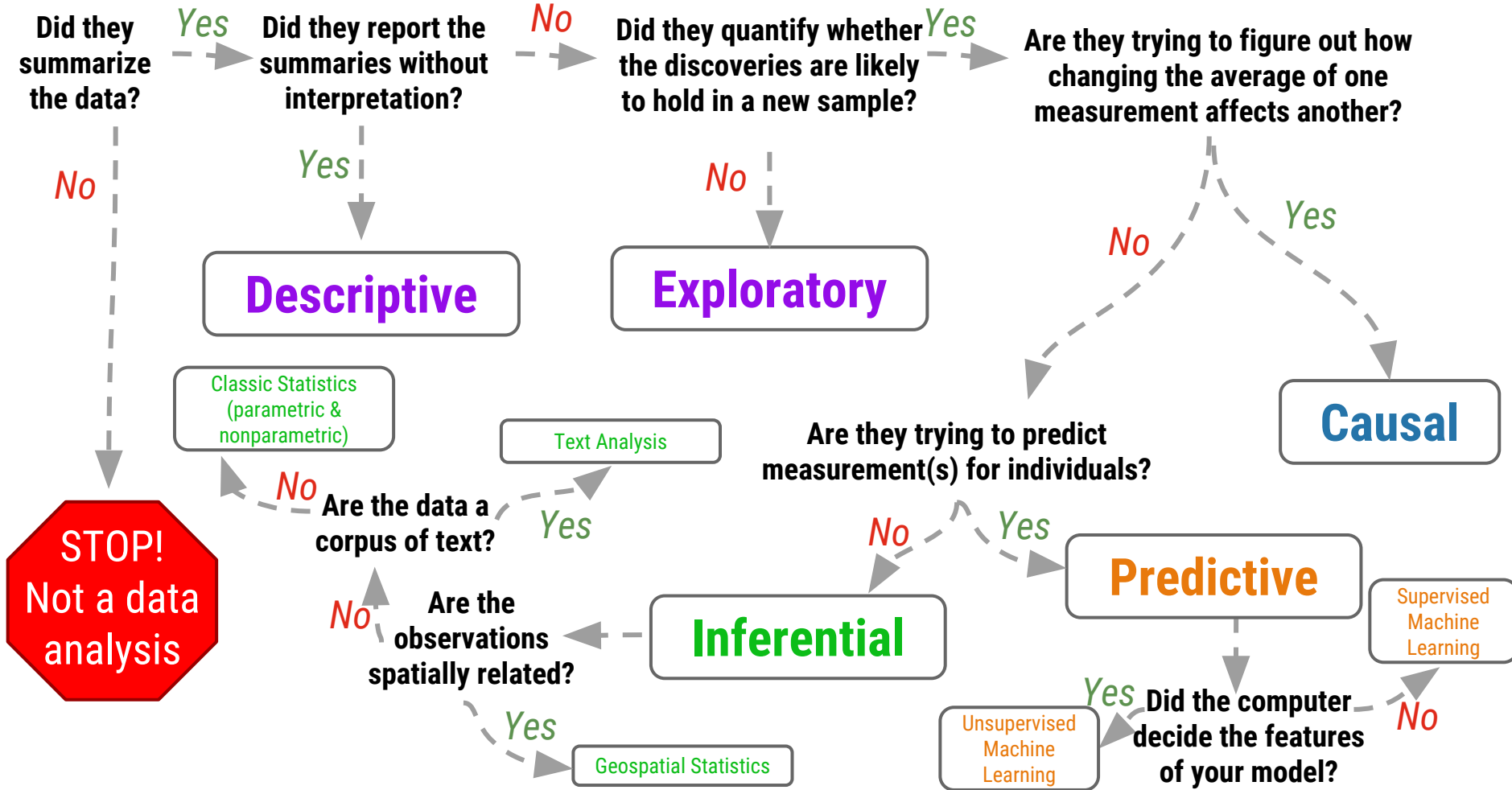
A EDA goes beyond providing summaries of the data

B EDA does not care about the shape of the data

C EDA only looks at single variables - not relationships

D EDA not used for communicating results

E EDA doesn't identify patterns in the data



Exploring Analyses

General question: What impacts politics in America?

Data Science question: Is there a relationship between the sentiment of political words in South Park and America's presidential approval rating?

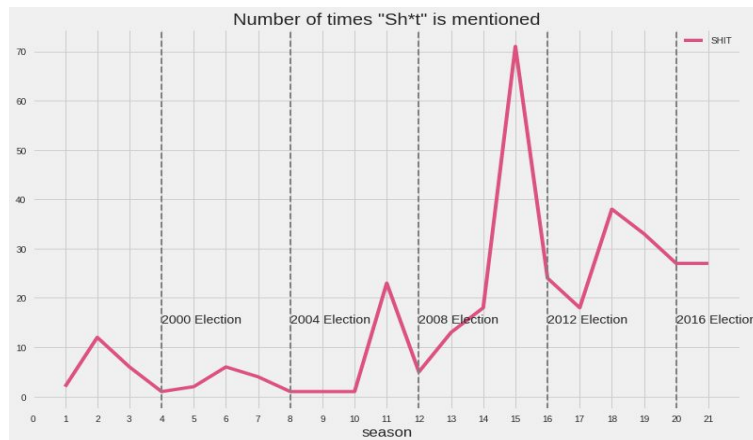
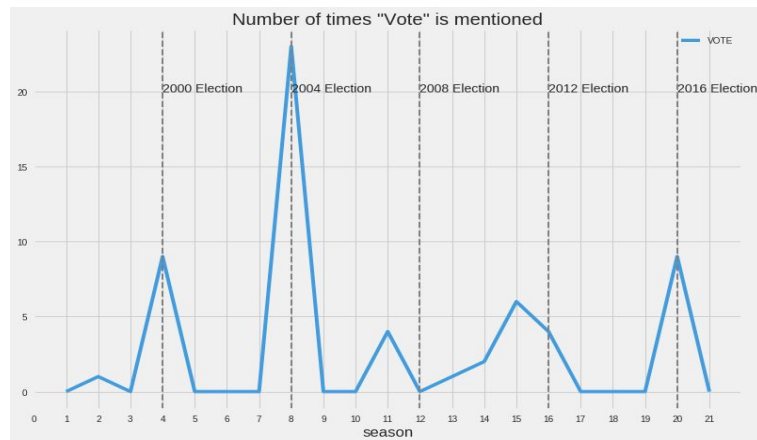
Descriptive

Exploratory

Inferential

Text Analysis

Classic Statistics
(parametric &
nonparametric)



General question: What gets too much attention in the news?

Data Science Question: Is there a relationship over time between cause of death terms in the *NYT*, *The Guardian*, and Google trends data relative to data from the CDC?

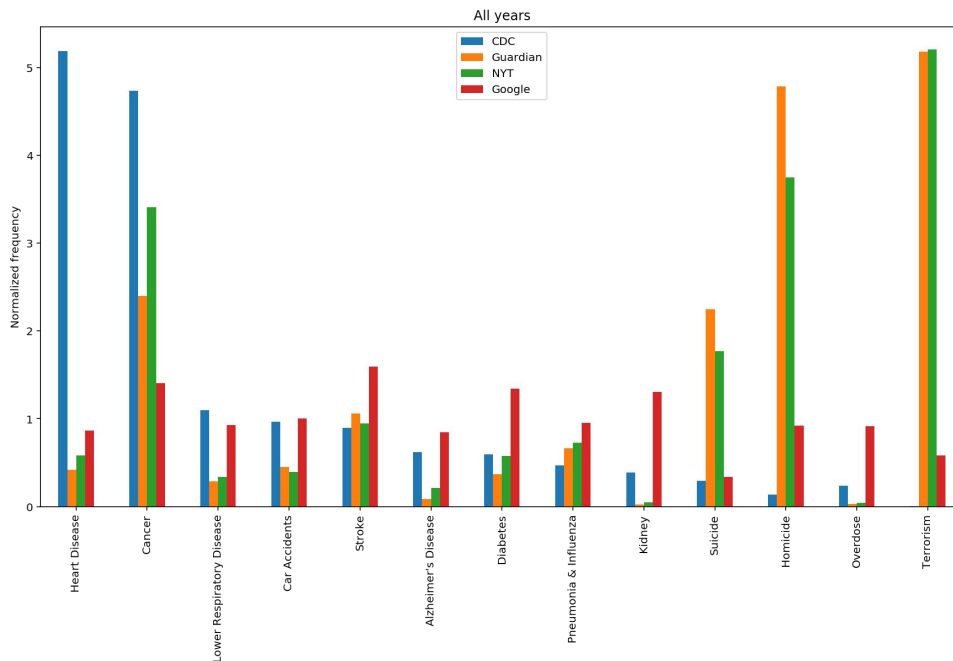
Descriptive

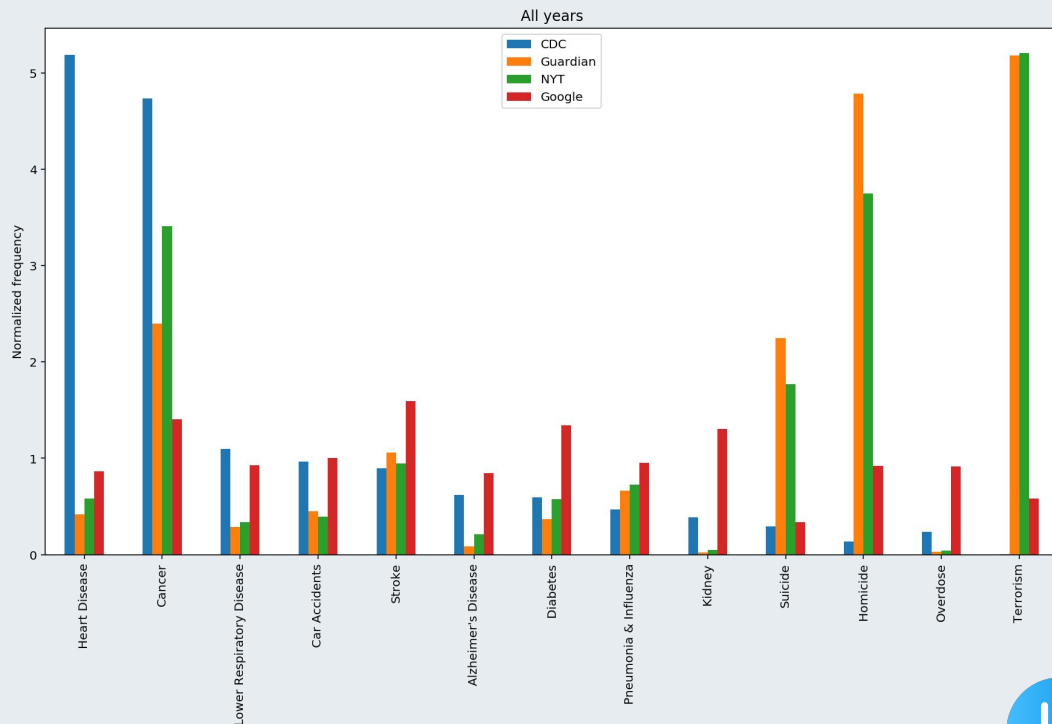
Exploratory

Inferential

Text Analysis

Classic Statistics
(parametric &
nonparametric)





What are positive aspects of this plot?

How would you improve this data display?



A when you've got some thoughts on how to improve
B if you're stuck and don't know how to improve

What is the relationship between the number of each genre of film produced per year and San Diego's crime rate?

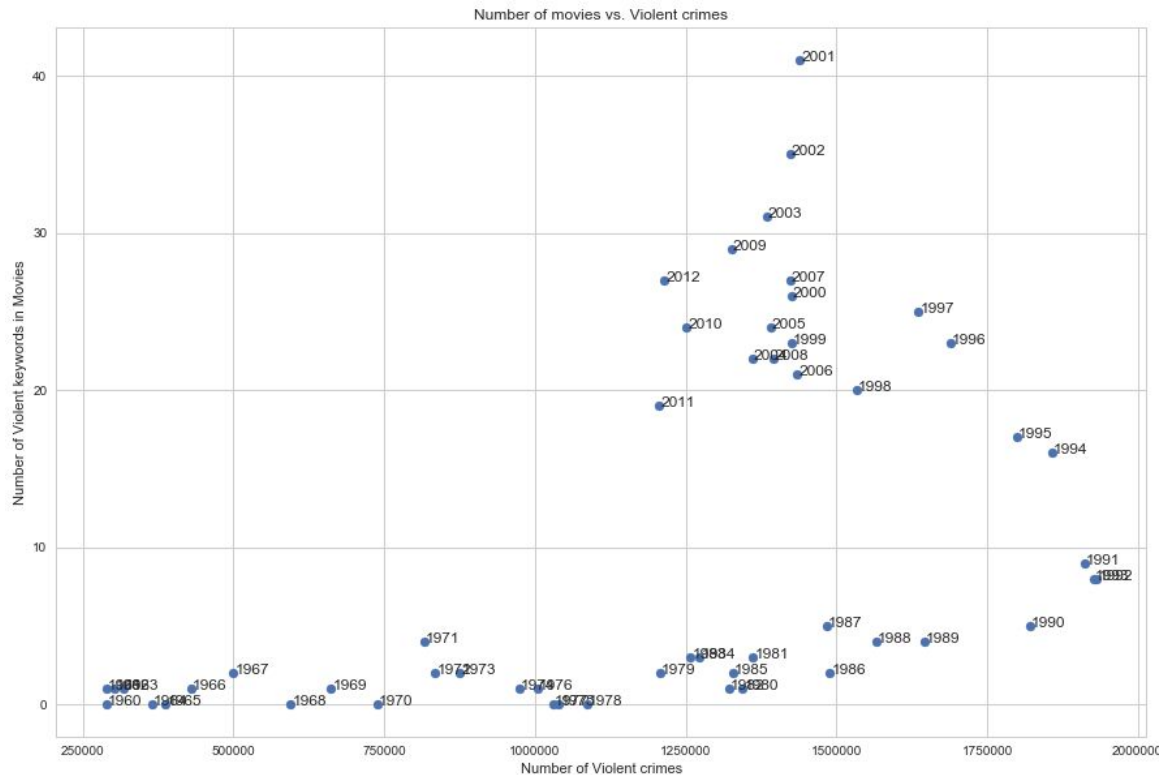
Descriptive

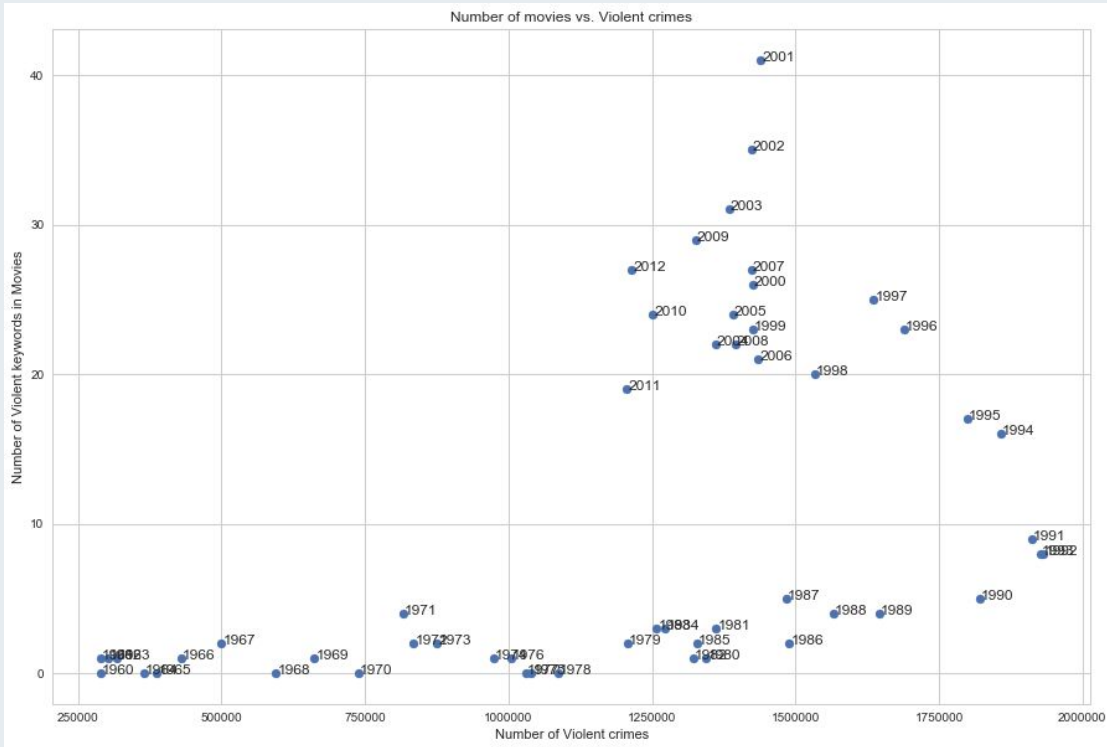
Exploratory

Inferential

Text Analysis

Classic Statistics
(parametric &
nonparametric)





What are positive aspects of this plot?

How would you improve this data display?



A when you've got some thoughts on how to improve
B if you're stuck and don't know how to improve

Descriptive

Exploratory

Predictive

Classification: Often we seek to assign a label to an item from a discrete set of possibilities.

Can we predict who will win next year's NCAA tournament? The Masters? The Super Bowl? The pennant? A game?

Can we predict the genre of a given movie (comedy, drama, or animation?) from just its script?

Descriptive

Exploratory

Predictive

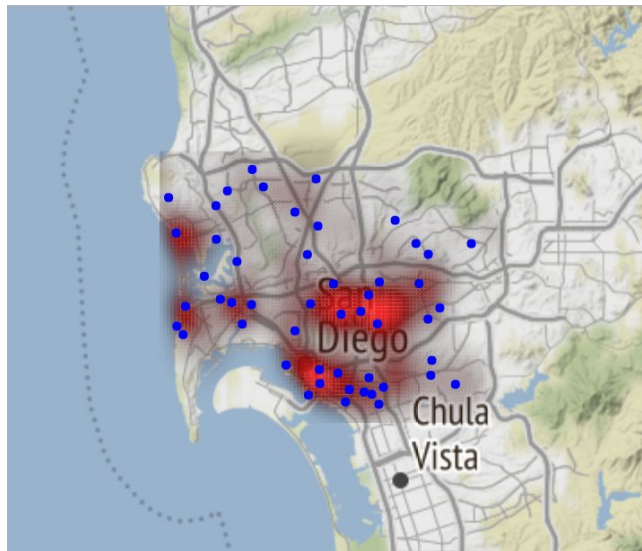
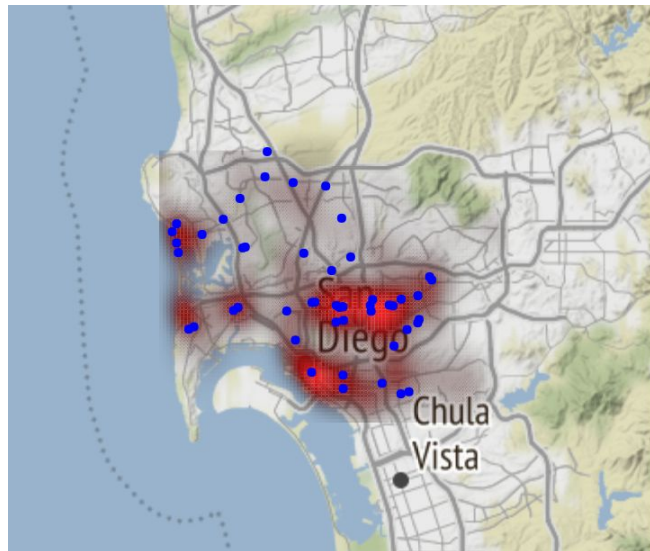
Regression: A way to forecast a given numerical quantity using other relevant features.

Can we predict someone's weight given other information?

How much snow will the East Coast get this year?

General question: Why isn't police response time always the same?

Data Science question: Where should police cars be stationed, accounting for crime levels and time of day, to make police response times equitable throughout San Diego?



Descriptive

Exploratory

Predictive

Inferential

In case of the total drought in California, how many desalination plant projects we need to supply residential use water for population who live in urban areas in California?

Descriptive

Exploratory

Predictive

