Descriptive and Exploratory Analysis

Jason G. Fleischer, Ph.D.
Asst. Teaching Professor
Department of Cognitive Science, UC San Diego
jfleischer@ucsd.edu

@jasongfleischer

https://jgfleischer.com

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**"*

A statistic:

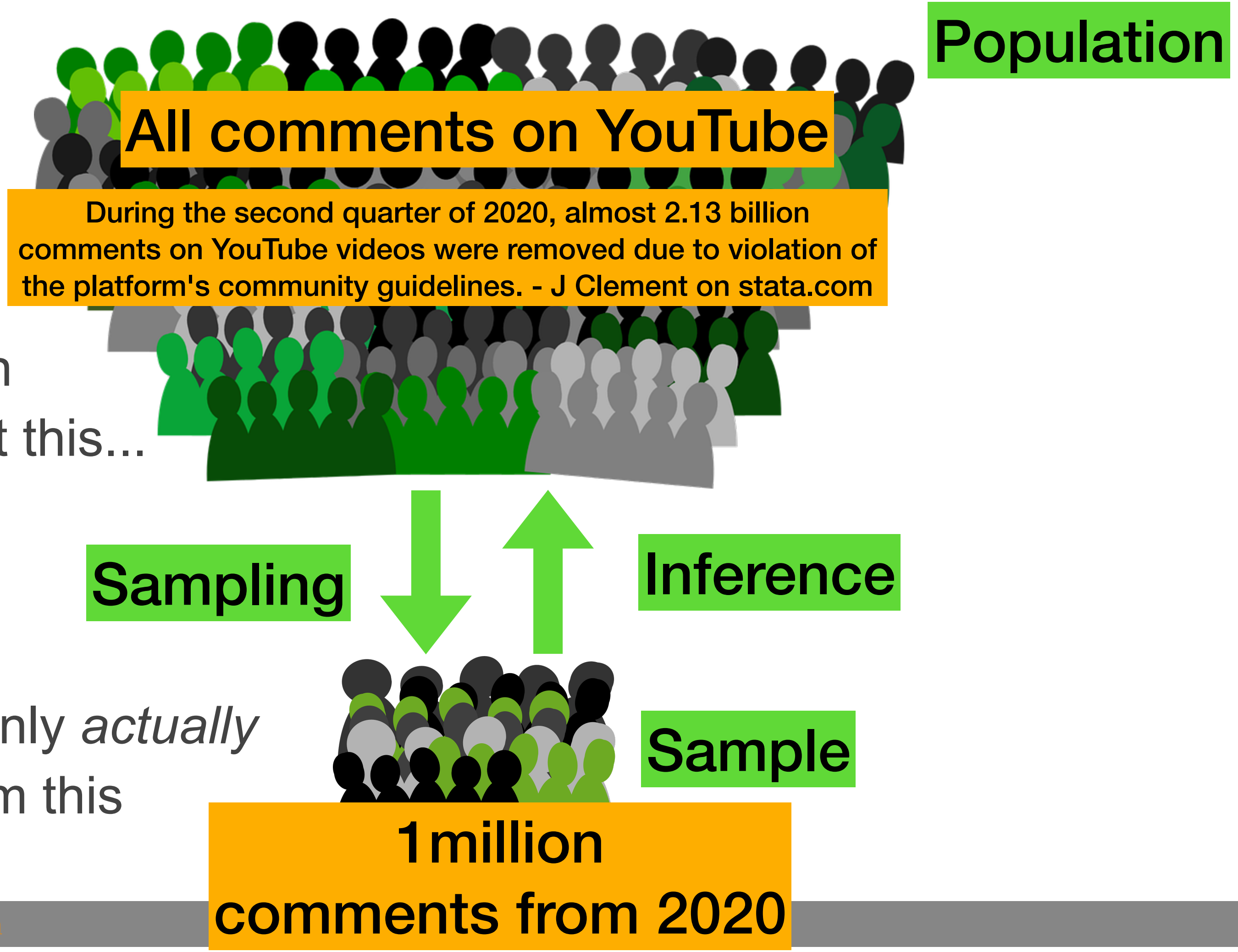
“*A quantity computed from a* *sample*”

statistic

"A quantity computed from a <u>sample</u>"

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

We want to learn something about this...



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

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.

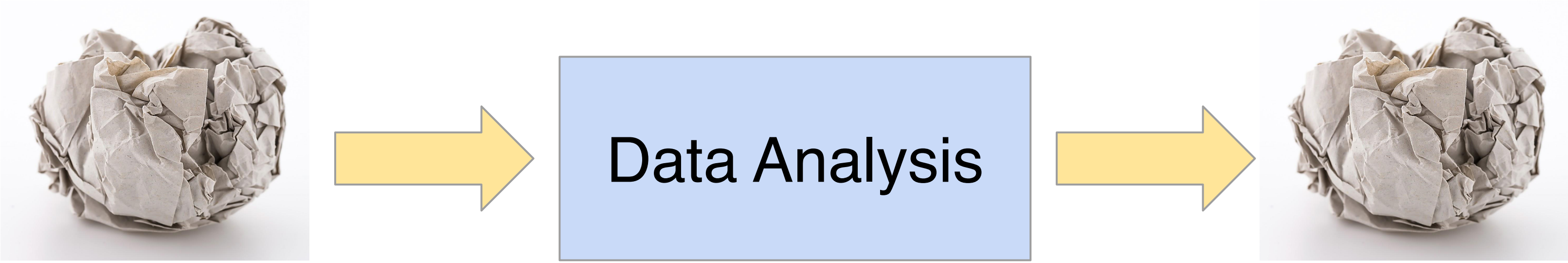
Examples of bad sampling:

- 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 <u>beginning</u> 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.



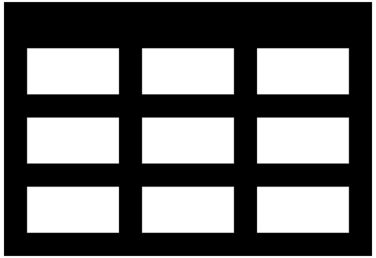
https://forms.gle/v9bn5kCUVLW5Robc6

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

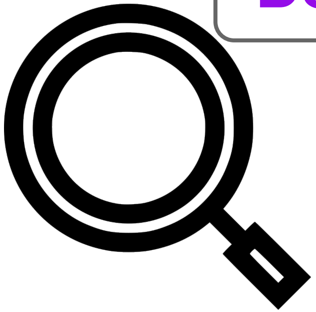
A B C D E

Descriptive

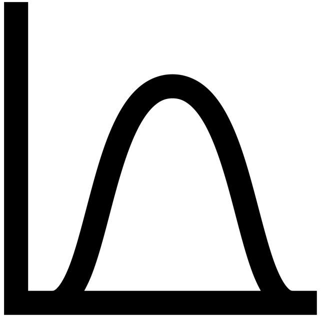
Descriptive Analysis



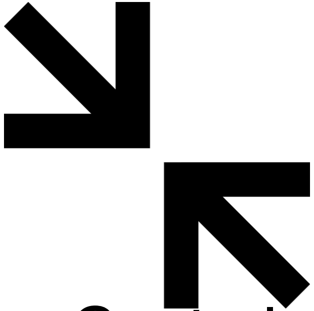
Size



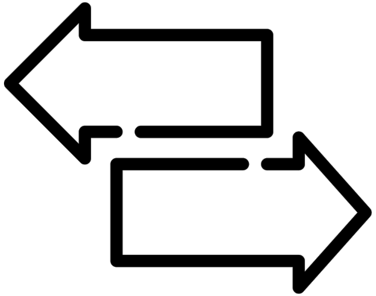
Missingness



Shape

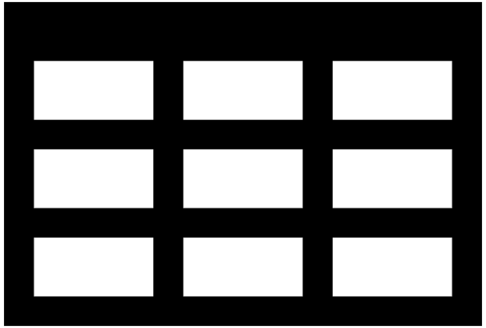


Central Tendency



Variability

Descriptive



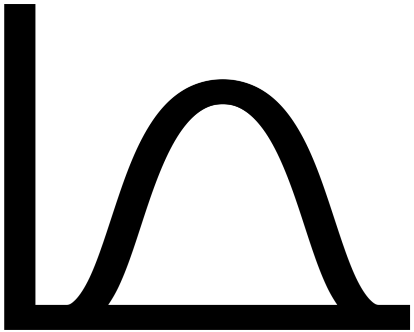
Size

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

Descriptive

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

Descriptive

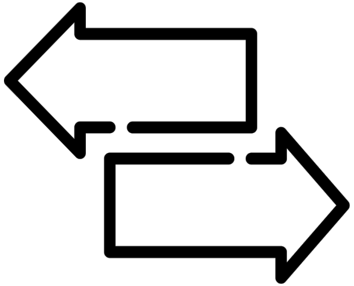
It’s critical to know the distribution of the variables in your dataset. Certain statistical approaches can only be used with certain distributions.

Descriptive

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

Descriptive



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

Descriptive Analyses are often included as "Table 1" in academic publications

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)
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.

Descriptive

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.

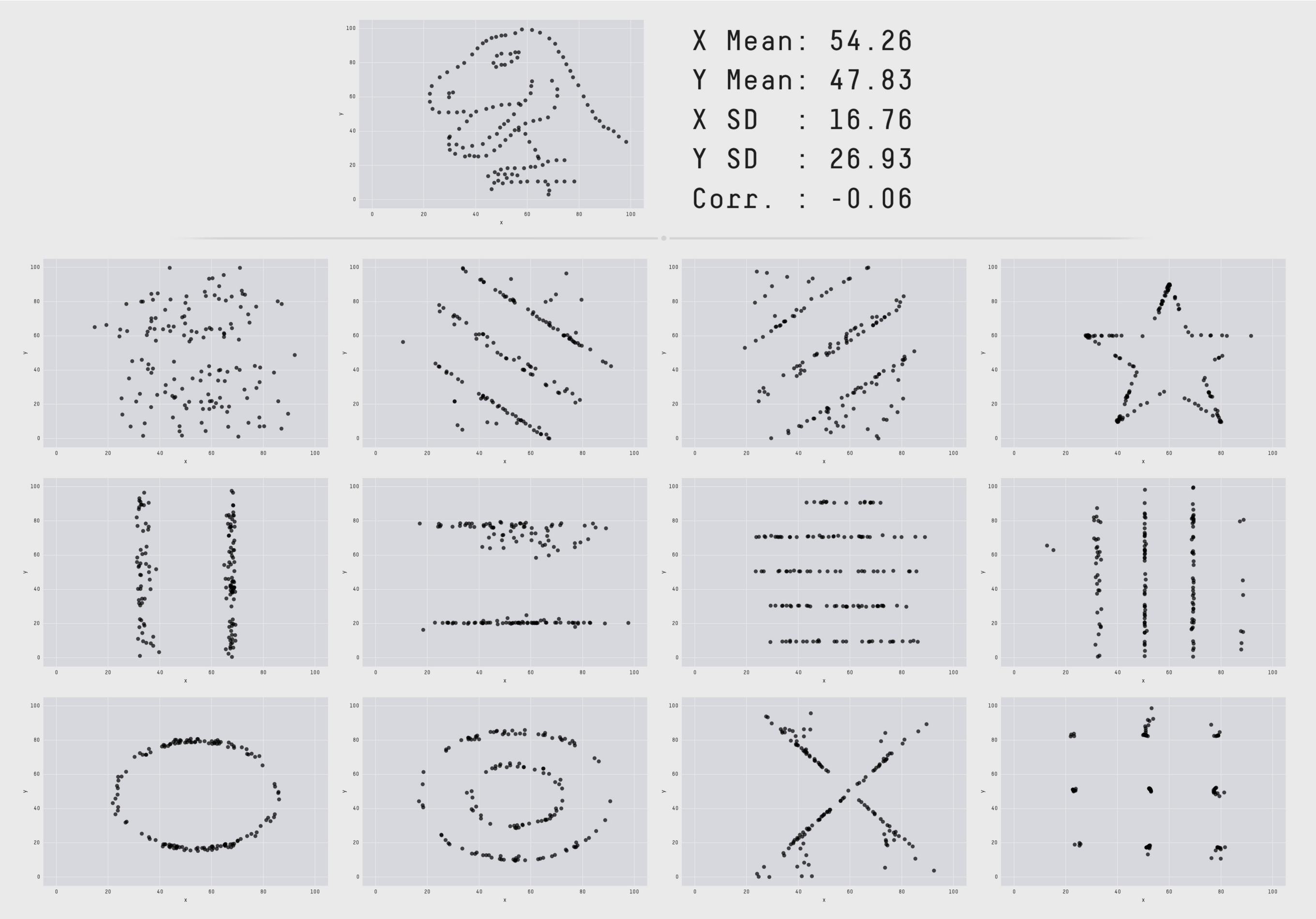
Size

Shape

Central tendency

variability

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



https://www.autodesk.com/research/publications/same-stats-different-graphs

A modern version of Anscombe's Quartet

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)

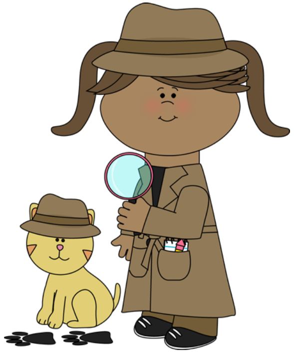
Descriptive Statistics & Summary

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

Yes statistics are summaries that throw away important detail, but human minds need the high level overview since we often struggle with the details.

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?*"

Exploratory

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

The dataset

You

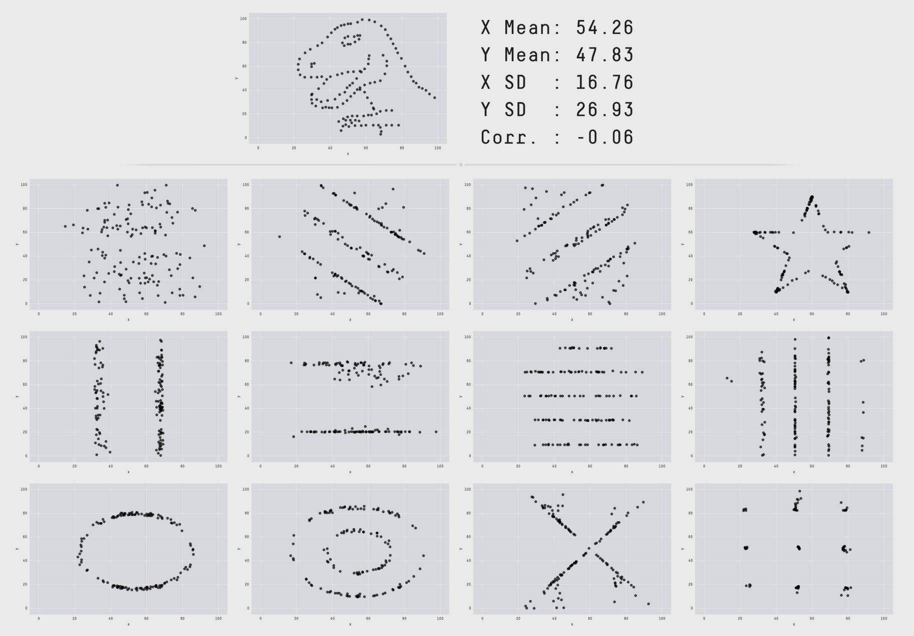
Exploratory

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

Start raw

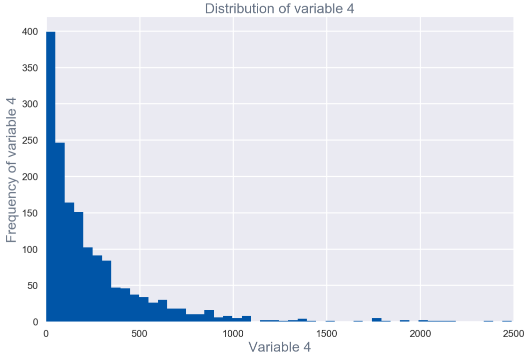
- Examine raw data in the most direct way you can reasonably do so
- View a random sample of the data
- Plots, especially subsets of variables and dimensionality reduction
- Helpful for seeing weirdness, missingness, outliers, min/max/typical values



https://www.autodesk.com/research/publications/same-stats-different-graphs

EDA Approaches to “Get a Feel for the Data”

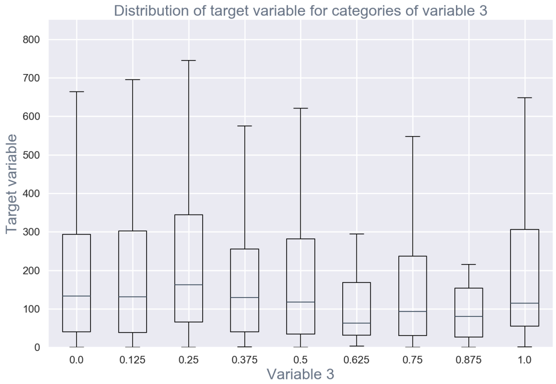
Understanding the relationship between variables in your dataset



Univariate

understanding a single variable

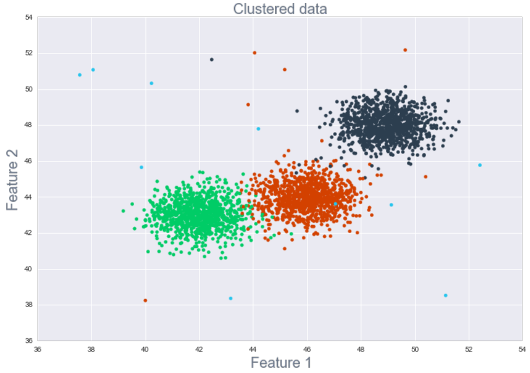
i.e.: histogram, densityplot, barplot



Bivariate

understanding relationship between 2 variables

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

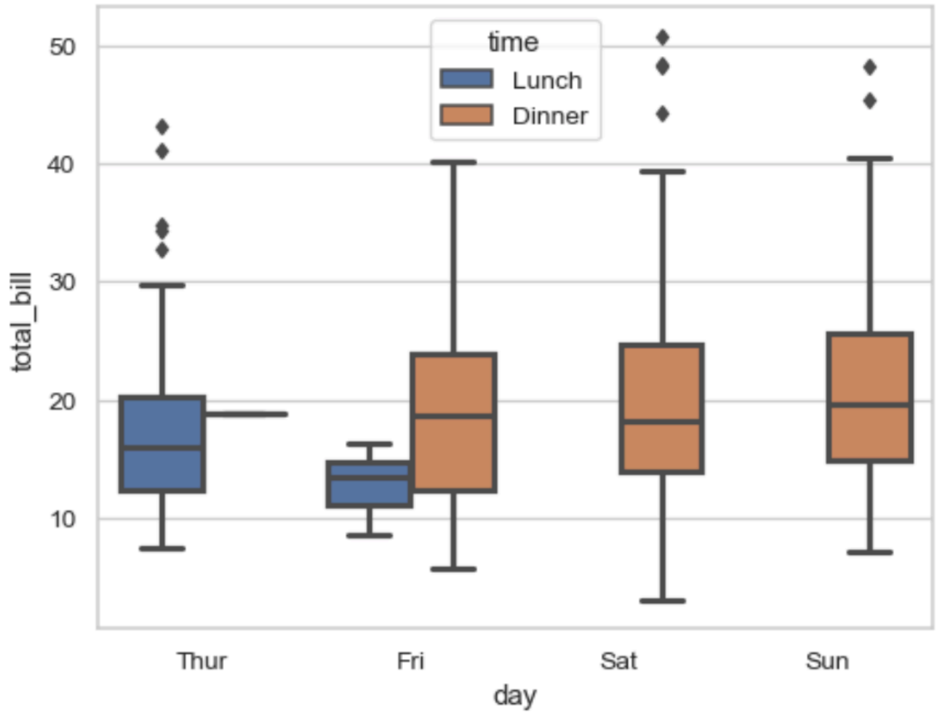


Dimensionality Reduction

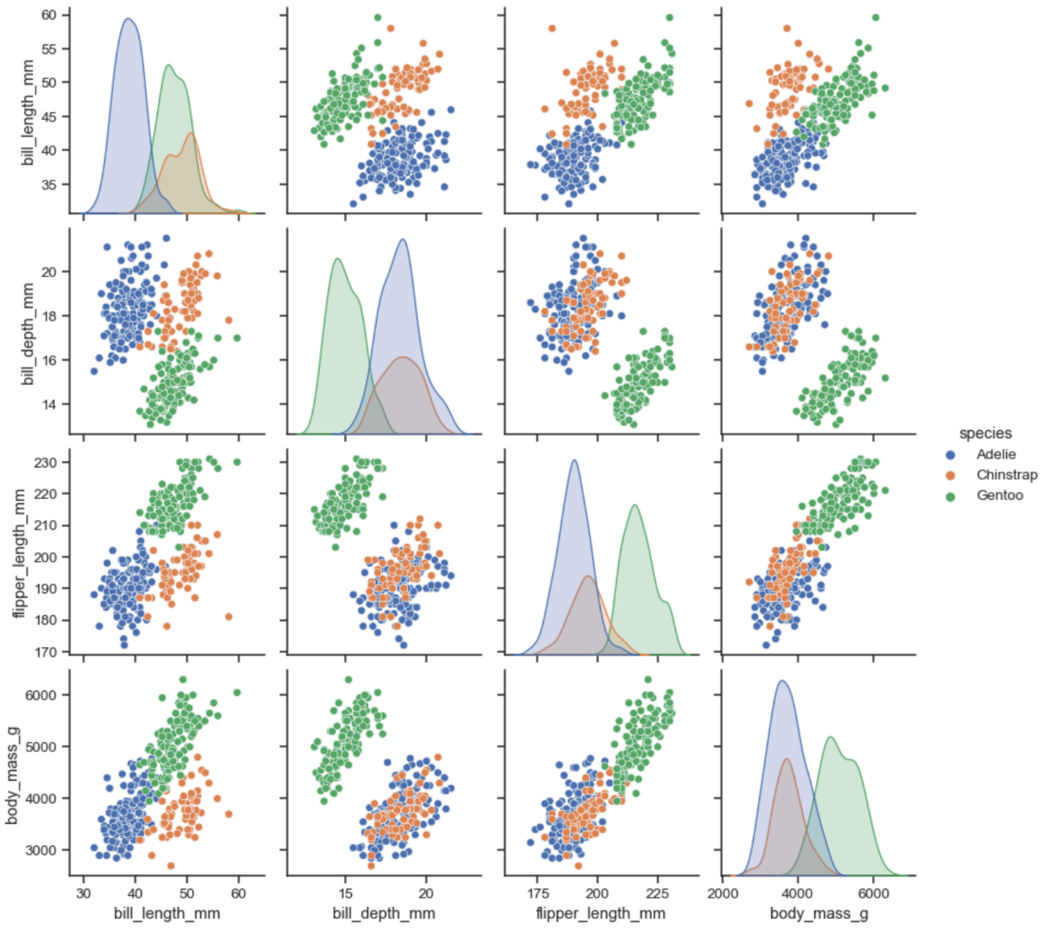
projecting high-D data into a lower-D space

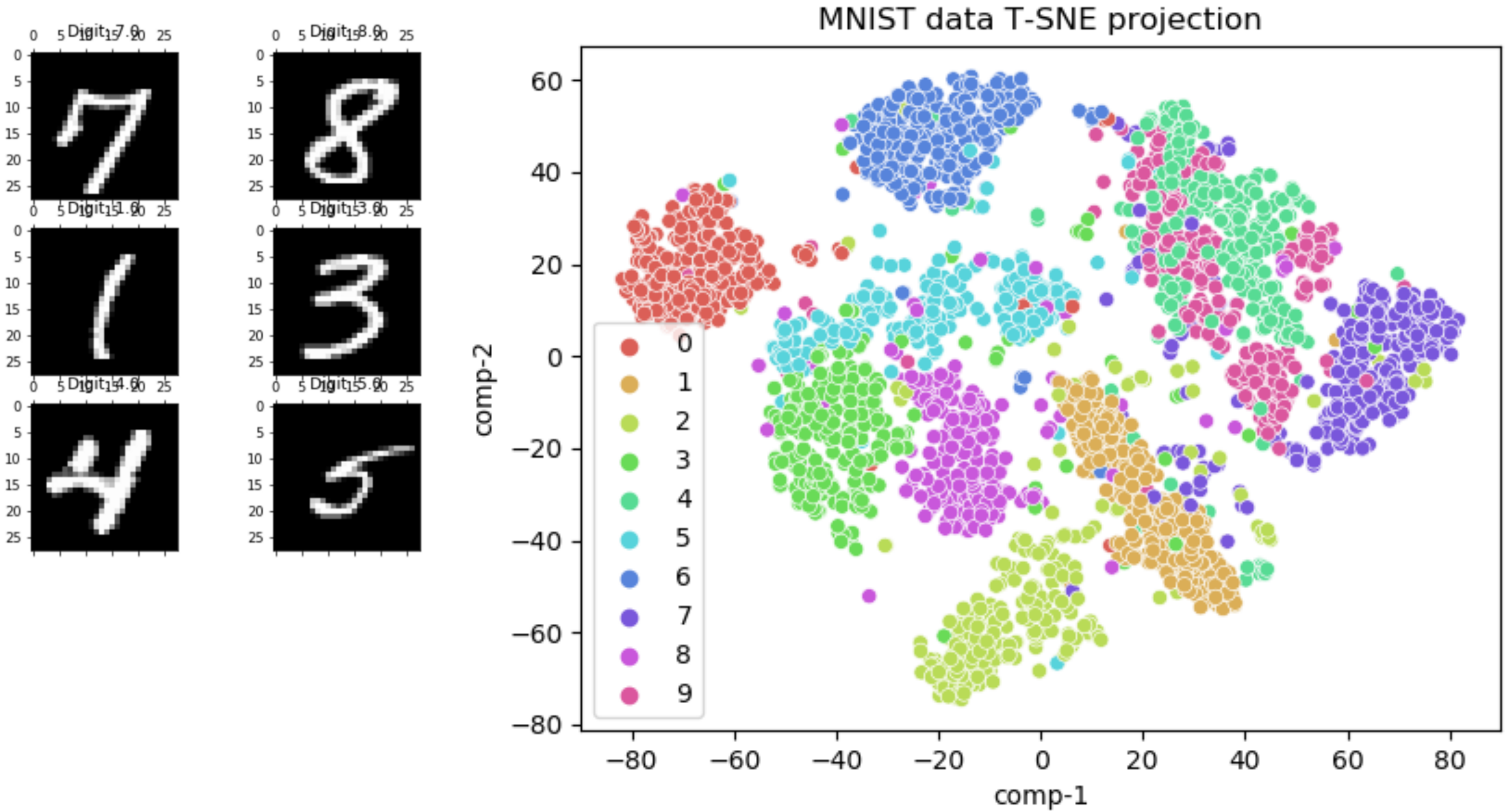
i.e.: PCA, ICA, Clustering

```
>>> ax = sns.boxplot(x="day", y="total_bill", hue="time",
...                  data=tips, linewidth=2.5)
```



```
sns.pairplot(penguins, hue="species")
```





MNIST data T-SNE projection
comp-1
comp-2
0
1
2
3
4
5
6
7
8
9