GENERAL ASSEMBLY

UNIT 01 - REVIEW

Abbas Chokor, Ph.D.

Staff Data Scientist, Seagate Technology

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS	~~What is Data Science~~	~~Lesson 1~~	
	~~Research Design and Pandas~~	~~Lesson 2~~	
	~~Statistics Fundamentals I~~	~~Lesson 3~~	
	~~Statistics Fundamentals II~~	~~Lesson 4~~	
	Flexible Class Session	Lesson 5	Today's Class
UNIT 2: FOUNDATIONS OF DATA MODELING	Introduction to Regression	Lesson 6	
	Evaluating Model Fit	Lesson 7	
	Introduction to Classification	Lesson 8	
	Introduction to Logistic Regression	Lesson 9	
	Communicating Logistic Regression Results	Lesson 10	
	Flexible Class Session	Lesson 11	
UNIT 3: DATA SCIENCE IN THE REAL WORLD	Decision Trees and Random Forests	Lesson 12	
	Natural Language Processing	Lesson 13	
	Dimensionality Reduction	Lesson 14	
	Time Series Data I	Lesson 15	
	Time Series Data II	Lesson 16	
	Database Technologies	Lesson 17	
	Where to Go Next	Lesson 18	
	Flexible Class Session	Lesson 19	
	Final Project Presentations	Lesson 20	

LAST CLASS

WHAT DID WE LEARN?

- Review of statistics and go over unit project 1
- Explain the difference between causation and correlation
- Test a hypothesis within a sample case study
- Validate your findings using statistical analysis (p-values, confidence intervals)

LAST CLASS

ANNOUNCEMENTS

- Happy Hour – Thanks for coming
- Fill your exit ticket!
- Any other questions?

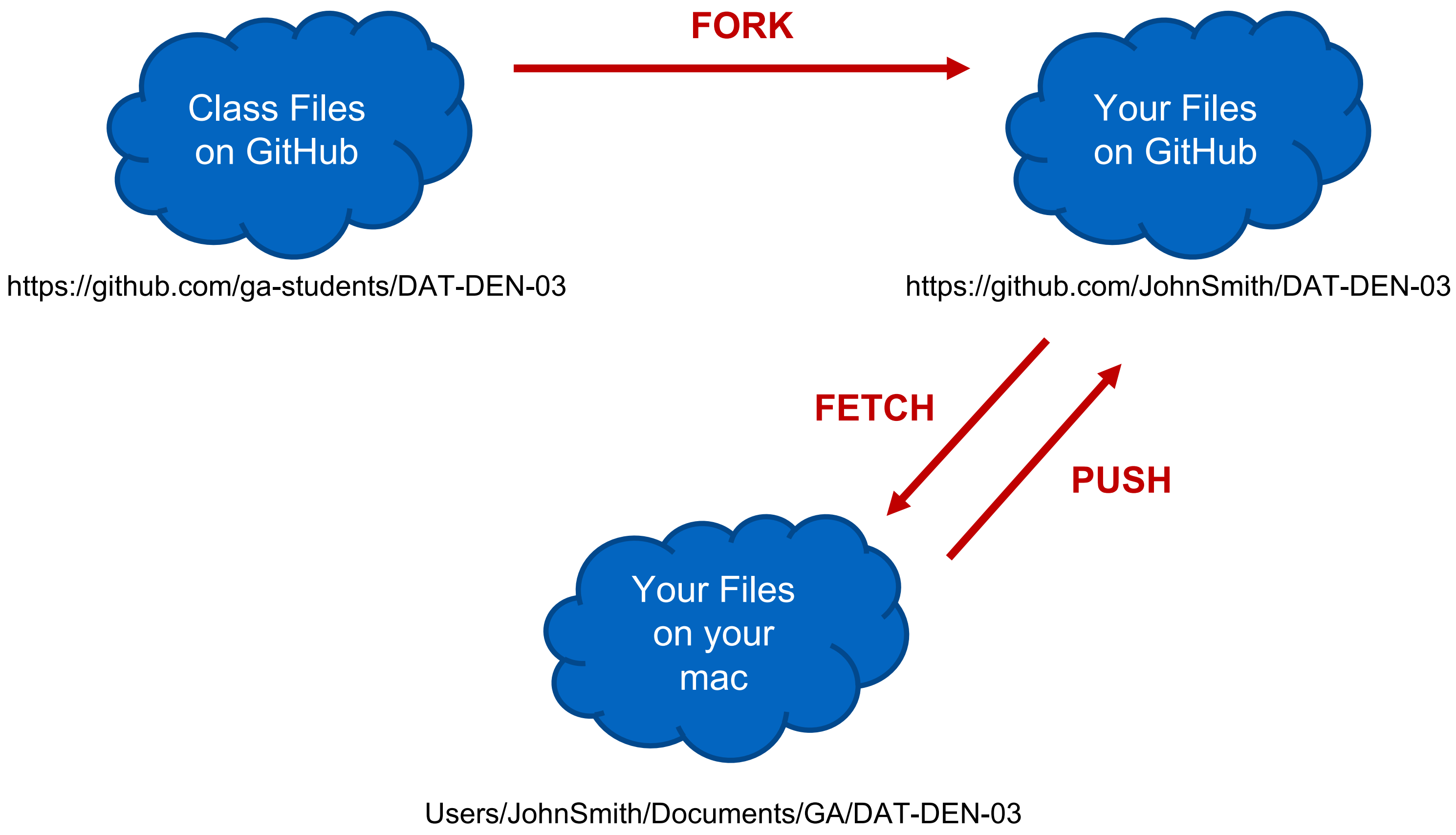
difference between histogram & density plot (googled it after class and understand now)

Others?

UNIT 01 - SUMMARY

Let’s Review What We Learned So Far…

HOW ARE WE GOING TO MANAGE OUR FILES?



FORK
Class Files on GitHub
https://github.com/ga-students/DAT-DEN-03
Your Files on GitHub
https://github.com/JohnSmith/DAT-DEN-03
FETCH
PUSH
Your Files on your mac
Users/JohnSmith/Documents/GA/DAT-DEN-03

HOW TO KEEP YOUR GITHUB UPDATED?

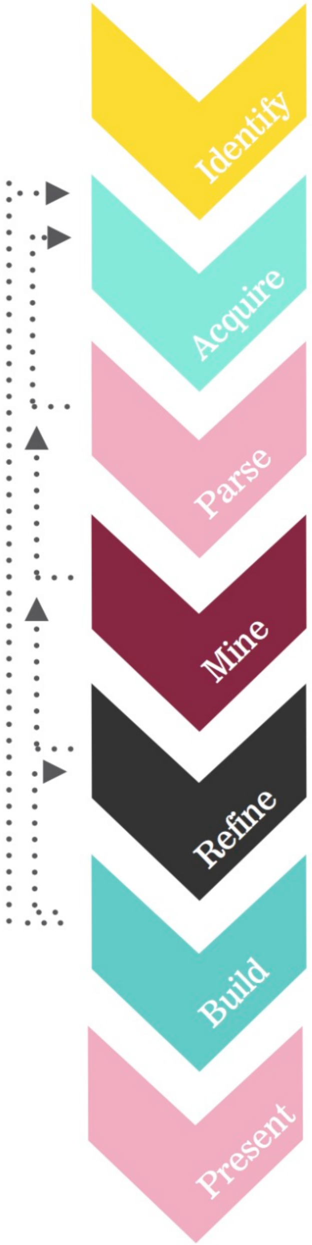
Synch to the class GitHub few hours after class using your Terminal.

```
git clone git@github.com/JohnSmith/DAT-DEN-03.git

cd /Users/665066/Documents/GitHub/DAT-DEN-03
git remote add upstream git://github.com/ga-students/DAT-DEN-03.git
git fetch upstream
git commit -m ”.” (if there is any change)
git pull upstream master
git push (to keep your online Github account synch with your local files)
```

Create and modify notebooks and python files…

LET’S REVIEW THE DATA SCIENCE WORKFLOW



IDENTIFY THE PROBLEM

□ Identify business/product objectives
□ Identify and hypothesize goals and criteria for success
□ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

□ Identify the “right” data set(s)
□ Import data and set up local or remote data structure
□ Determine most appropriate tools to work with data

PARSE THE DATA

□ Read any documentation provided with the data
□ Perform exploratory data analysis
□ Verify the quality of the data

MINE THE DATA

□ Determine sampling methodology and sample data
□ Format, clean, slice, and combine data in Python
□ Create necessary derived columns from the data (new data)

REFINE THE DATA

□ Identify trends and outliers
□ Apply descriptive and inferential statistics
□ Document and transform data

BUILD A DATA MODEL

□ Select appropriate model
□ Build model
□ Evaluate and refine model

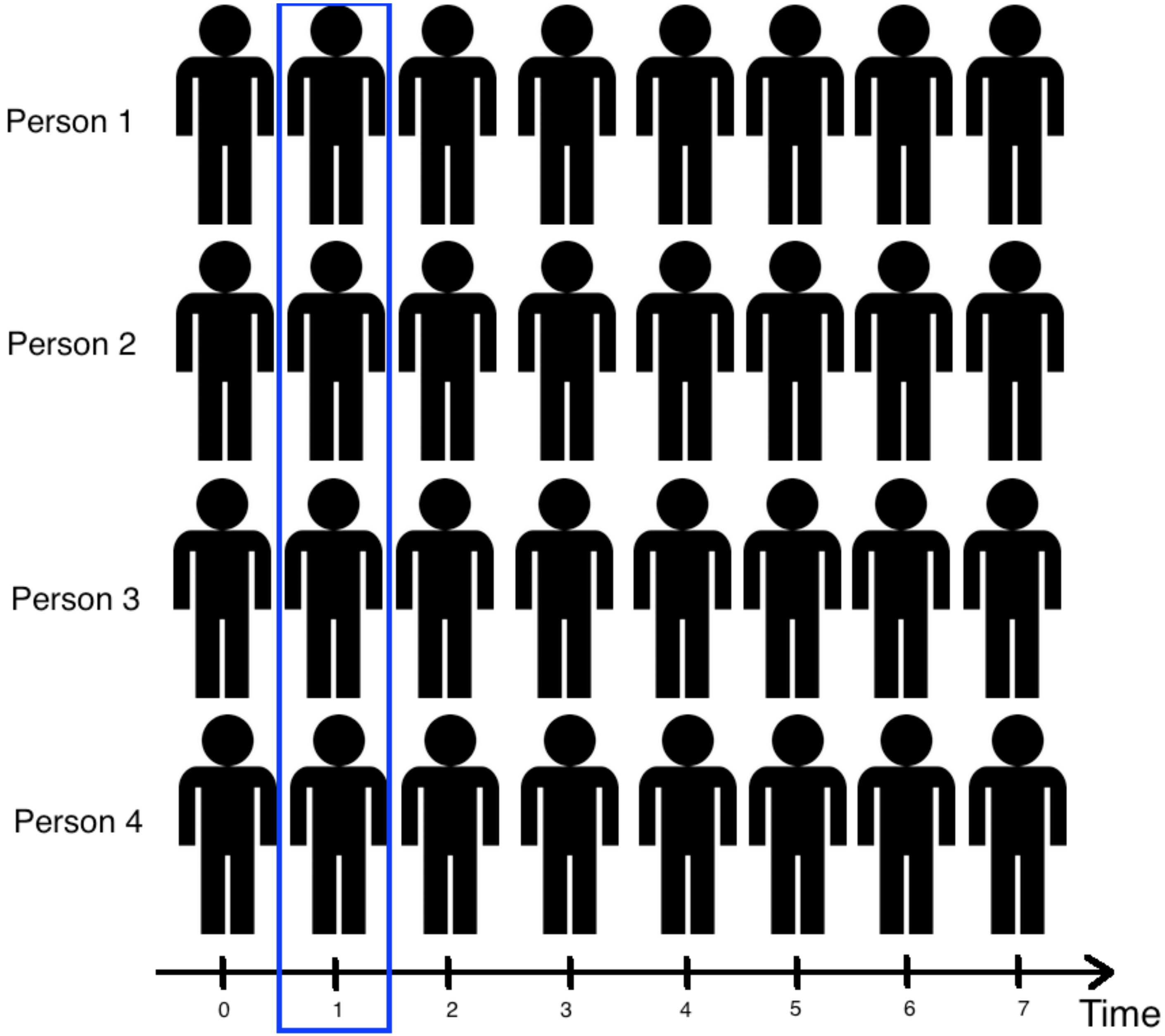
PRESENT THE RESULTS

□ Summarize findings with narrative, storytelling techniques
□ Present limitations and assumptions of your analysis
□ Identify follow up problems and questions for future analysis

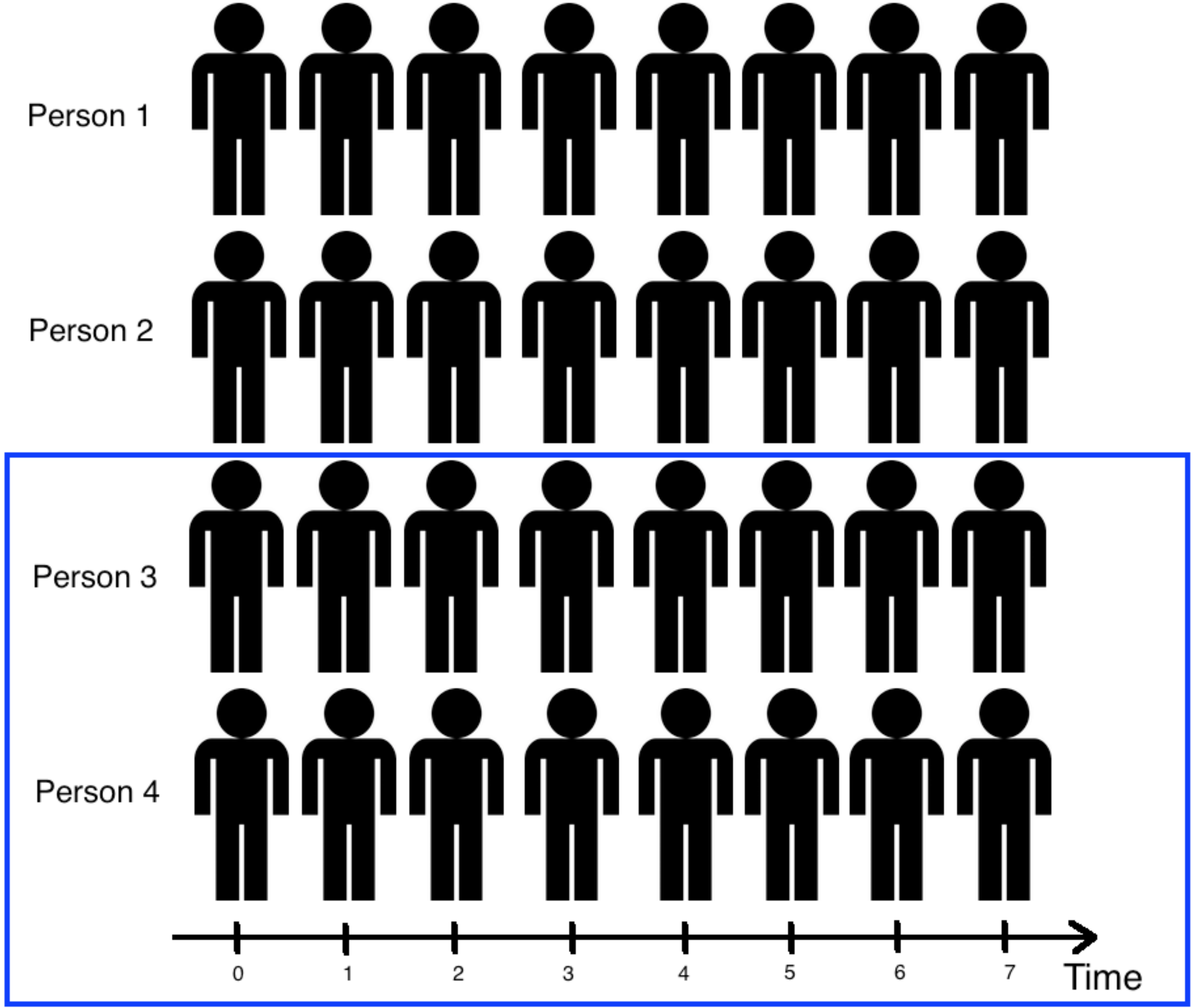
WHAT IS A GOOD QUESTION? SMART

- **S**pecific: The dataset and key variables are clearly defined.
- **M**easurable: The type of analysis and major assumptions are articulated.
- **A**ttainable: The question you are asking is feasible for your dataset and is not likely to be biased.
- **R**eproducible: Another person (or future you) can read and understand exactly how your analysis is performed.
- **T**ime-bound: You clearly state the time period and population for which this analysis will pertain.

CROSS-SECTIONAL DATA



TIME SERIES/LONGITUDINAL DATA

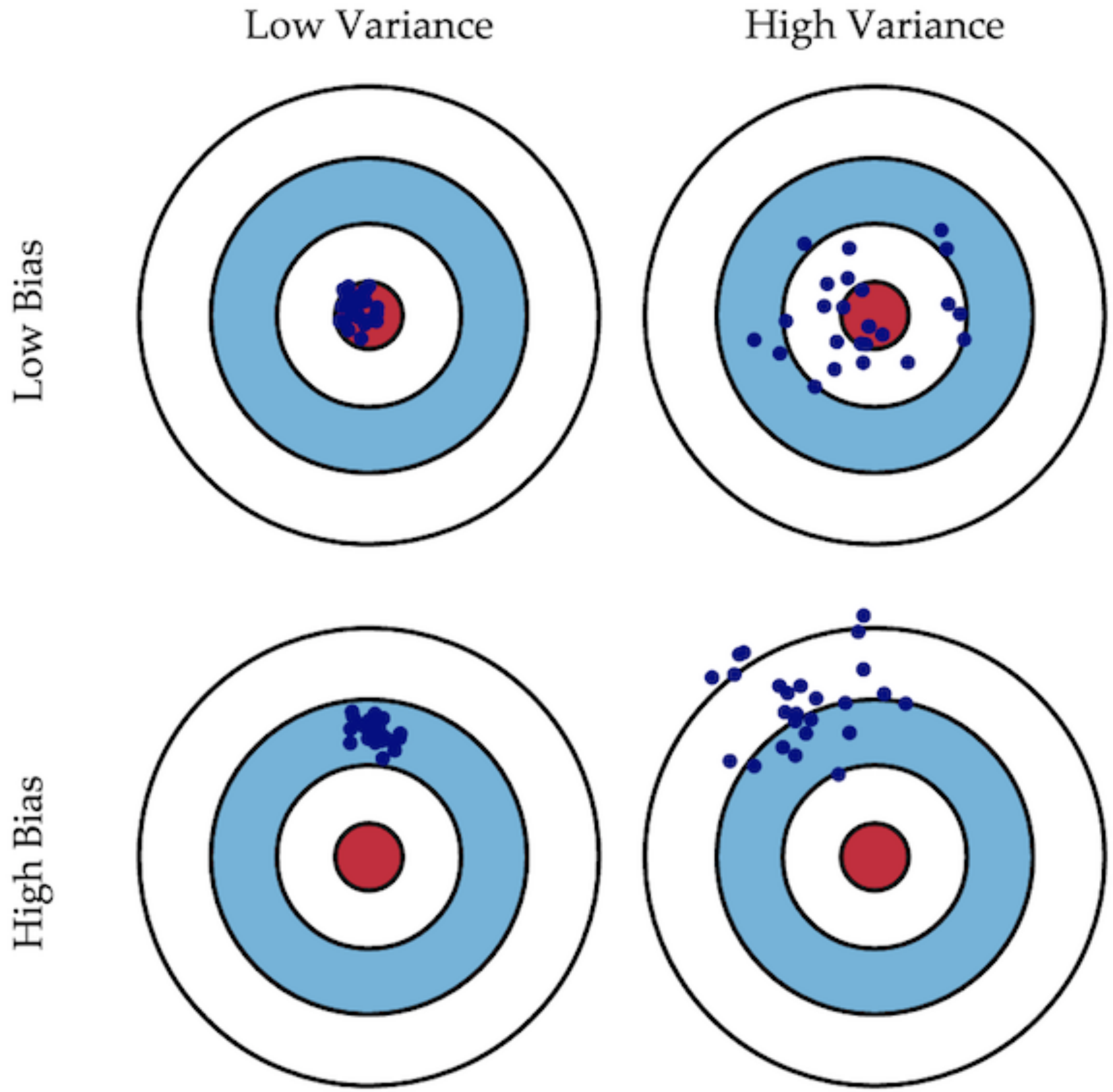


CODEALONG PART 1: BASIC STATS

- We can use Pandas to calculate the mean, median, mode, min, and max.

```
Methods available include:
.min() - Compute minimum value
.max() - Compute maximum value
.mean() - Compute mean value
.median() - Compute median value
.mode() - Compute mode value
.count() - Count the number of observations
```

BIAS VS. VARIANCE



CODEALONG PART 3: STANDARD DEVIATION & VARIANCE

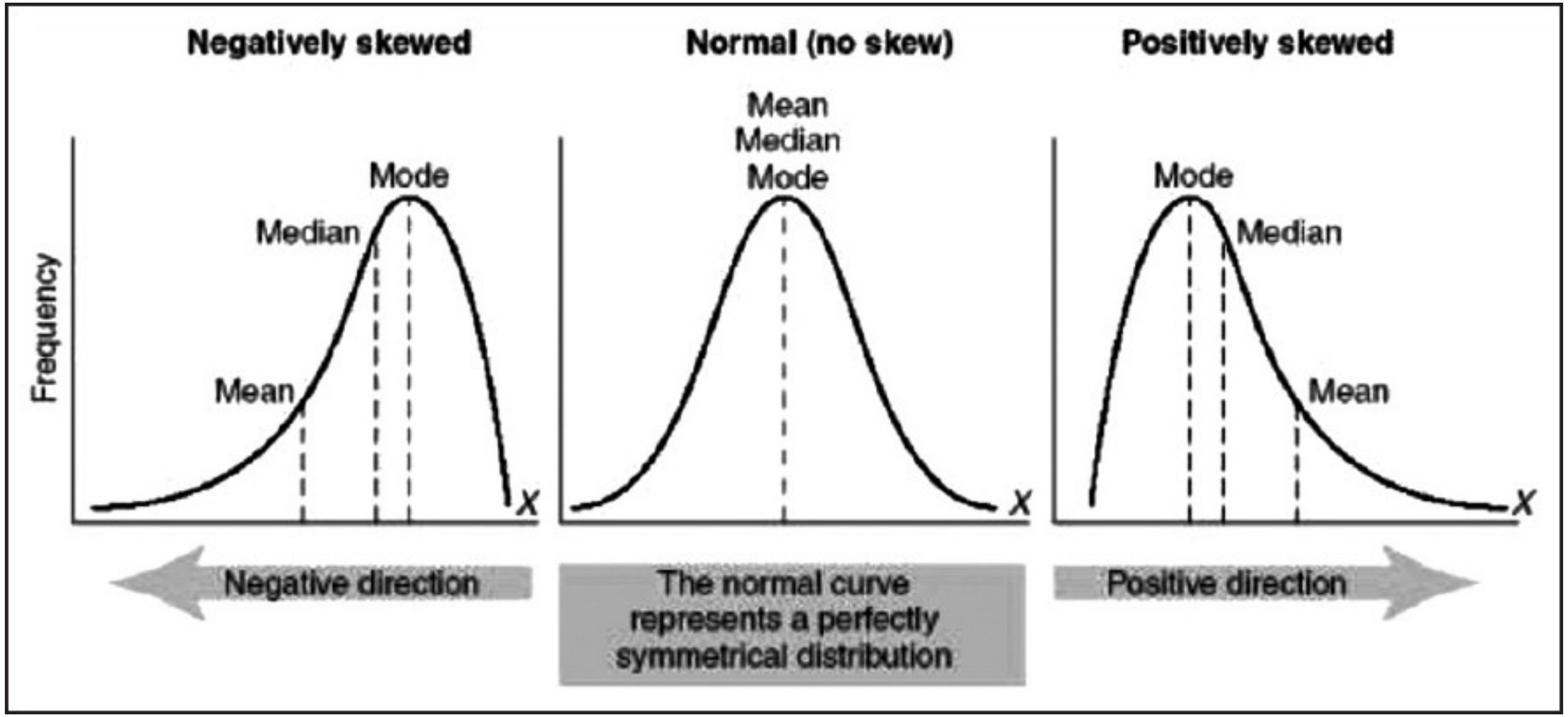
- You can calculate variance and standard deviation easily in Pandas.

```
Methods include:
.std() - Compute Standard Deviation
.var() - Compute variance
.describe() - short cut that prints out count, mean, std, min,
quartiles, max
```

SKEWNESS

- Skewness is a measure of the asymmetry of the distribution of a random variable about its mean.

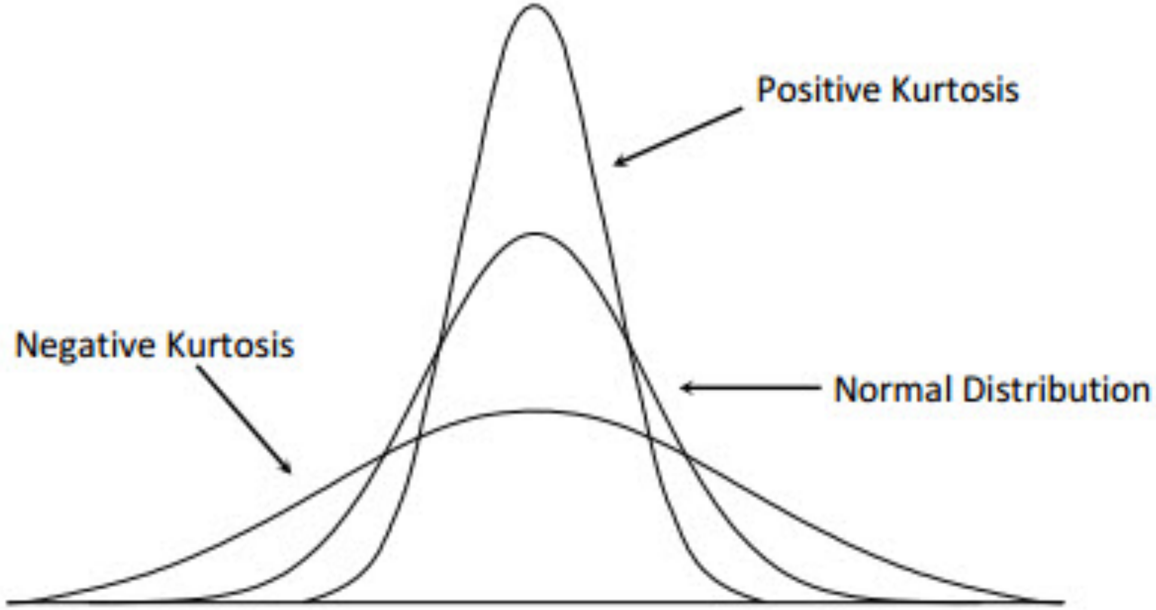
- Skewness can be positive or negative, or even undefined.



KURTOSIS

- Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution.

- Datasets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails.



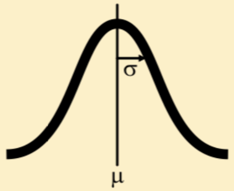
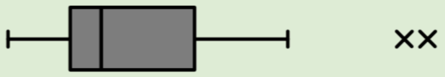
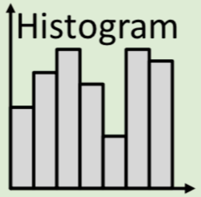
CLASS/DUMMY VARIABLES

- Step 1: Select a reference category. We'll choose `rural` as our reference category.

- Step 2: Convert the values `urban`, `suburban`, and `urban` into a numeric representation that does not imply order.

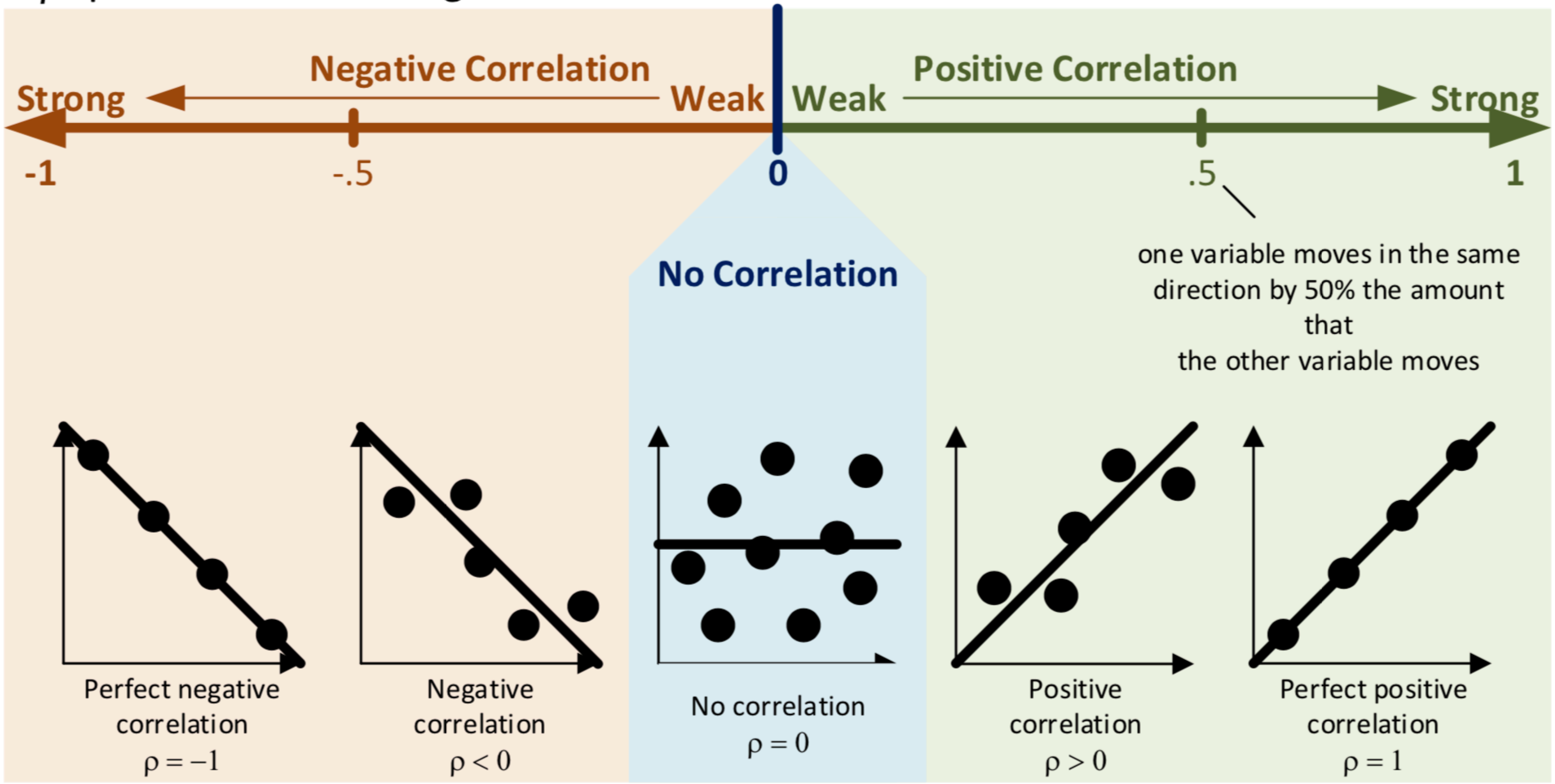
- Step 3: Create two new variables: `area_urban` and `area_suburban`.

STATS SUMMARY

Measure of Centrality	Mean	Median	Mode
Measurement Scales	Interval - Ratio	Interval - Ratio	Nominal - Ratio
• In the dataset?	☹	😐	☺
• Easy of compute	☺	😐	☹
• Resistant to outliers?	☹	☺	☺
Measure of Dispersion	☺ (Variance, Standard Deviation)	☺ (Interquartile Range)	☹
Extensive used in mathematical models?	☺	☹	☹
Graphical Methods		Boxplot 	

CORRELATION

ρ quantifies the strength and direction of movements of two random variables



PYTHON & PANDAS

Measure of Centrality	.mean()	.median()	.mode()

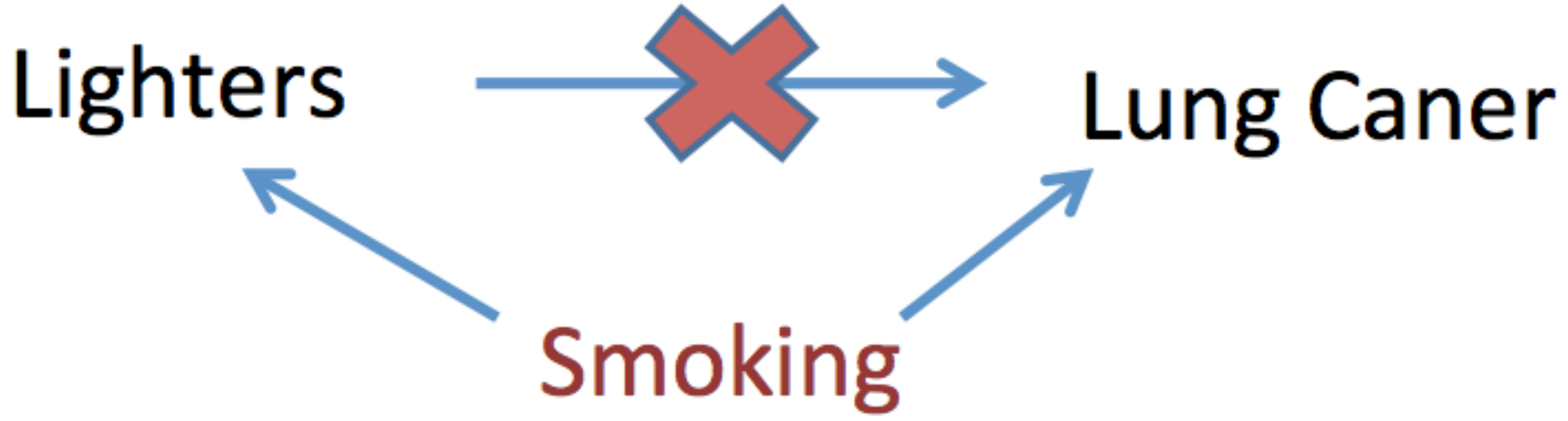
Measure of Dispersion	.var(), .std()	.min(), .max()
		.quantile()	

Summary	.describe()		
Graphical Methods		.plot(kind = 'box')	.plot(kind = 'hist')
Correlation Matrix	.corr()		
Scatter plot	*DataFrame*.plot(kind = 'scatter', x = '*SerieName*', y = '*SerieName*')		
Scatter matrix	pd.tools.plotting.scatter_matrix(*DataFrame*)		

.columns, .set_index(), .drop()	len(), .count(), .sum(), .unique() .value_counts(), .isnull(), .notnul(), .dropna()	np.sort(), .apply()

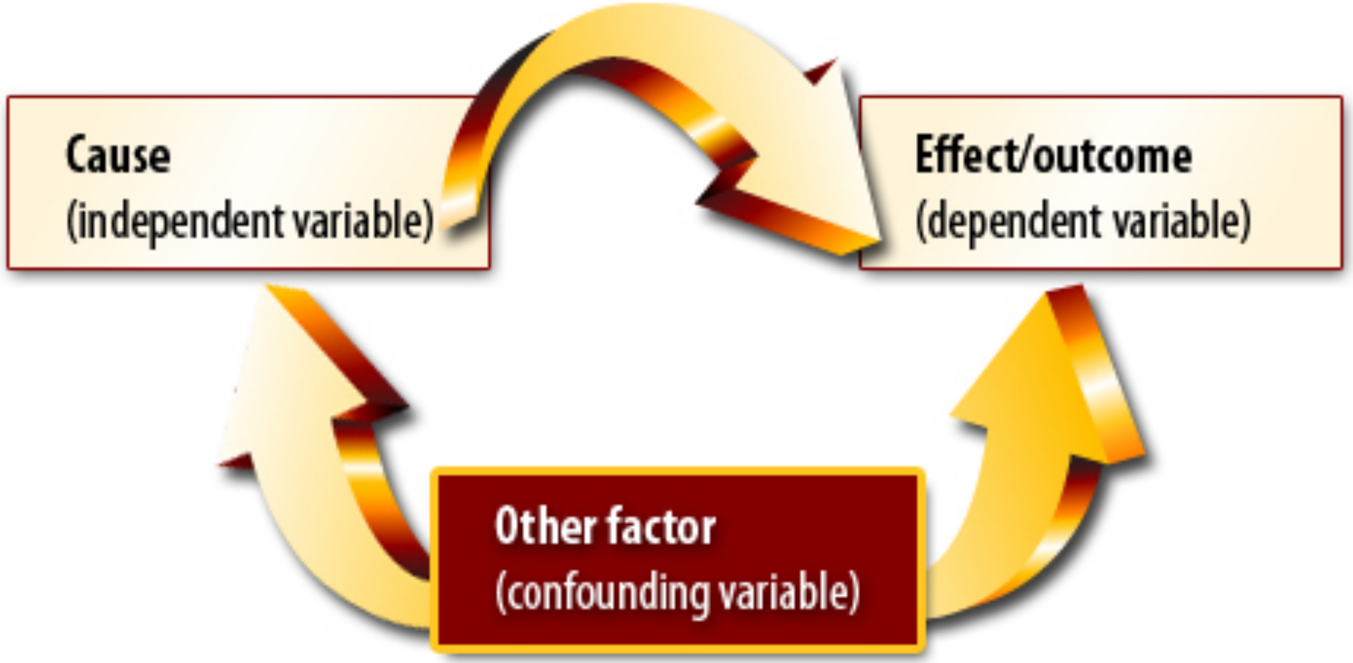
CONFOUNDING

- No!



CONFOUNDING

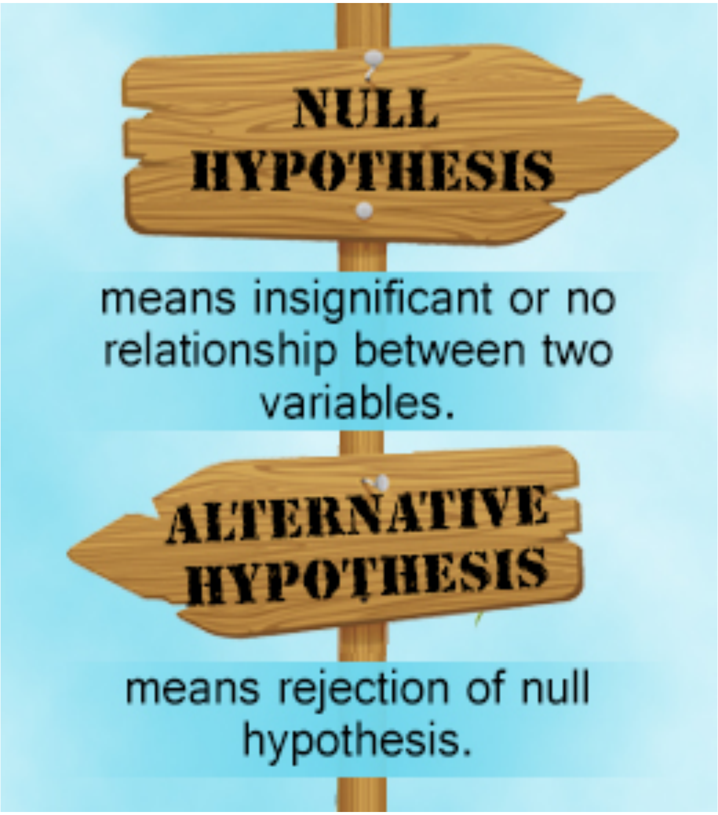
- Confounding variables often hide the true association between causes and outcomes.



In statistics, a confounder is a variable that influences both the dependent variable and independent variable causing a spurious association.

HYPOTHESIS TESTING STEPS

- First, you need a hypothesis to test, referred to as the *null hypothesis*. The opposite of this would be the *alternative hypothesis*.


NULL HYPOTHESIS
means insignificant or no relationship between two variables.
ALTERNATIVE HYPOTHESIS
means rejection of null hypothesis.

HYPOTHESIS TESTING STEPS

- For example, if we want to test the relationship between gender and sales, we may have the following hypotheses.

- Null hypothesis: There is no relationship between Gender and Sales.

- Alternative hypothesis: There is a relationship between Gender and Sales.

- Once you have your hypotheses, you can check whether the data supports rejecting the null hypothesis or failing to reject the hypothesis.

VALIDATE YOUR FINDINGS

- We know how to carry out a hypothesis test, but how do we tell if the association we found is *statistically significant*?

- *Statistical significance* is the likelihood that a result or relationship is caused by something other than random chance.

- Statistical hypothesis testing is traditionally employed to determine if a result is statistically significant or not.

VALIDATE YOUR FINDINGS

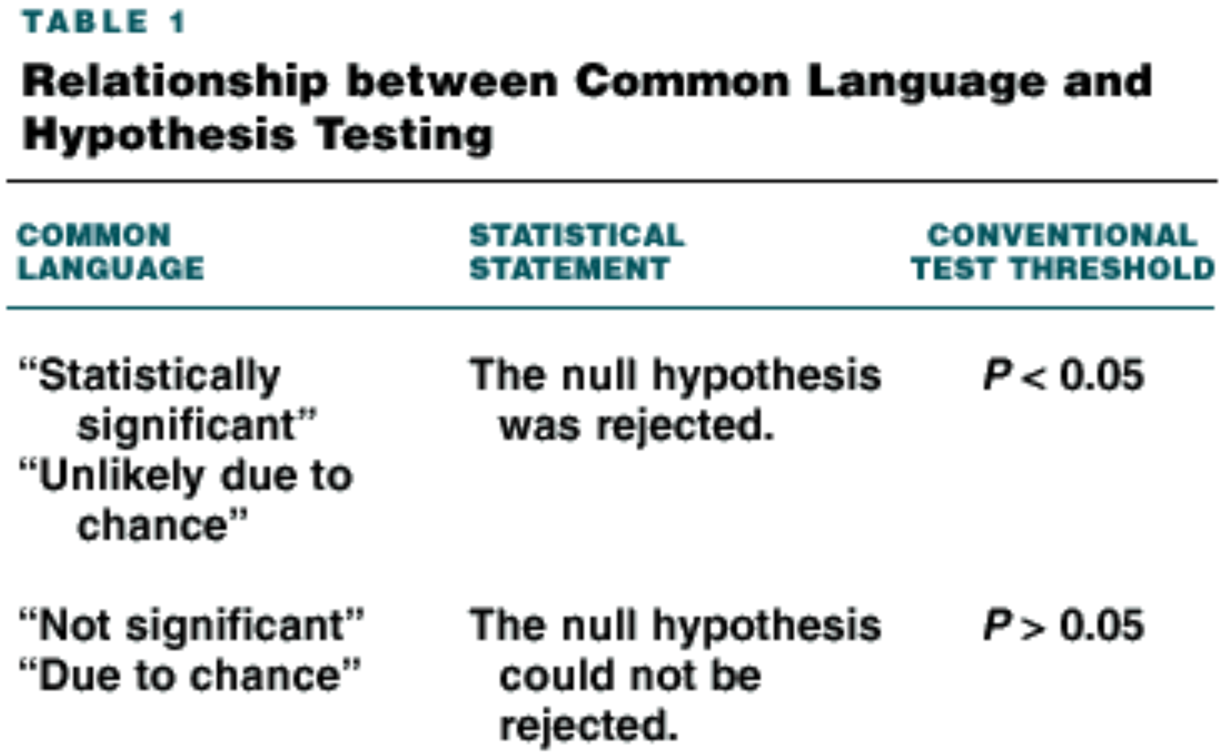
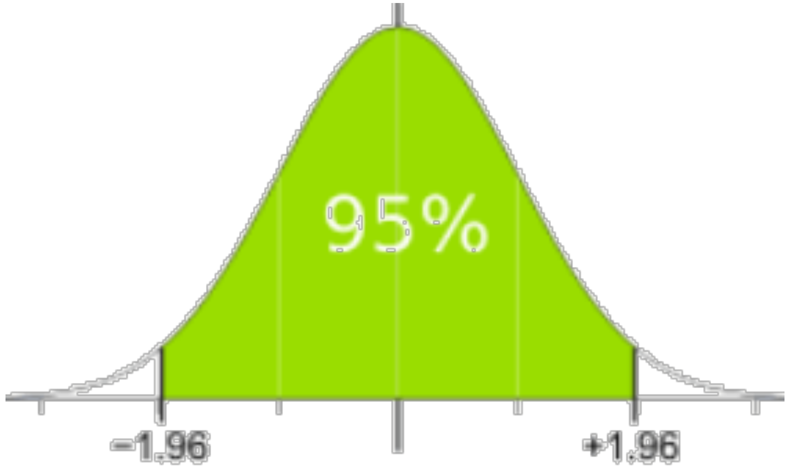


TABLE 1

Relationship between Common Language and Hypothesis Testing

COMMON LANGUAGE	STATISTICAL STATEMENT	CONVENTIONAL TEST THRESHOLD
"Statistically significant" "Unlikely due to chance"	The null hypothesis was rejected.	$P < 0.05$
"Not significant" "Due to chance"	The null hypothesis could not be rejected.	$P > 0.05$

If we repeat this study 100 times, our point of estimate would lie in that range 95 times.



VALIDATE YOUR FINDINGS

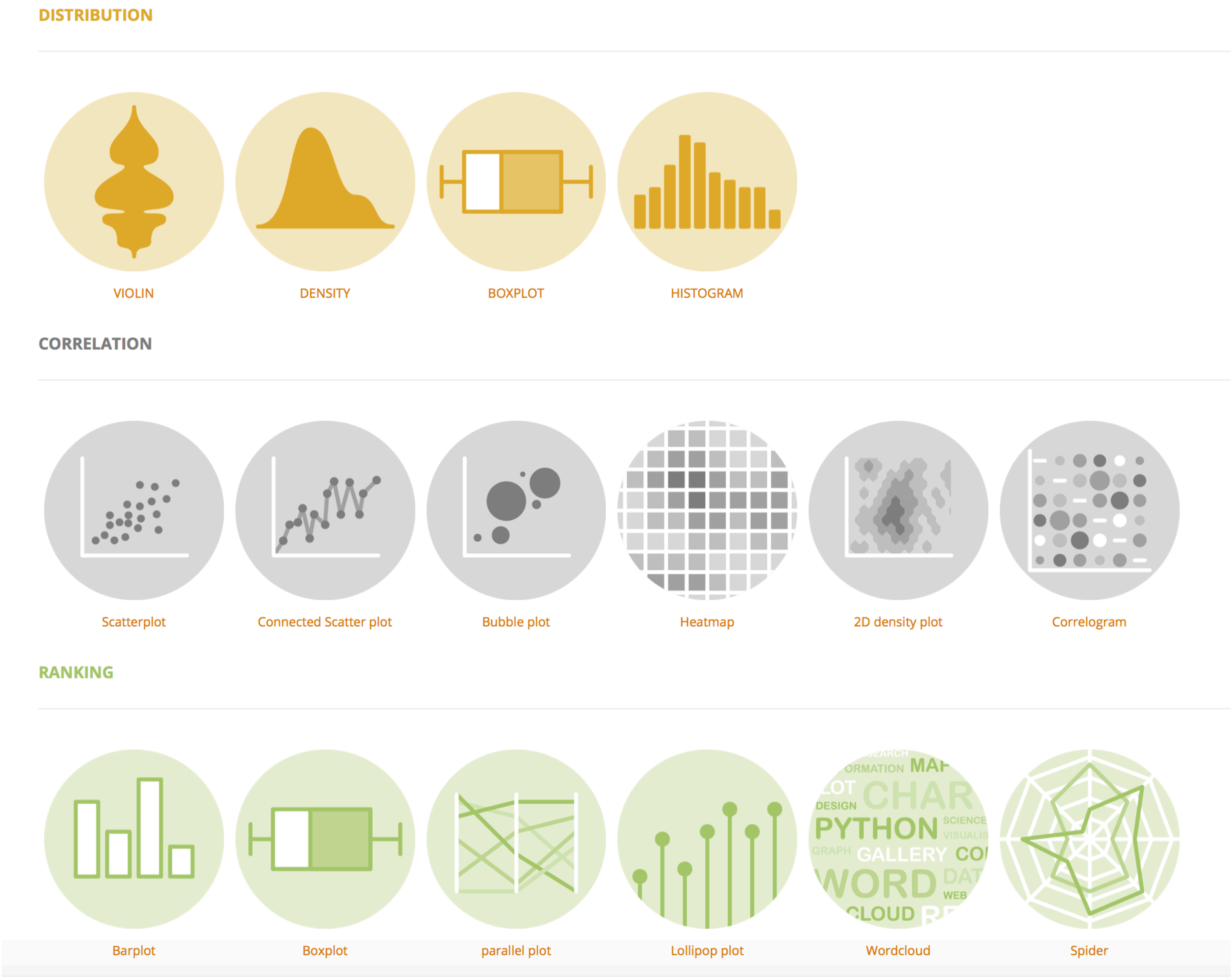
|t-value|	p-value	$1 - \alpha$ Confidence Interval $([\mu_0 - \cdot\ \sigma, \mu_0 + \cdot\ \sigma])$	H_0 / H_a	Conclusion
$\geq \cdot$	$\leq \alpha$	μ_0 is outside	Found evidence that $\mu \neq \mu_0$: Reject H_0	$\mu \neq \mu_0$
$< \cdot$	$> \alpha$	μ_0 is inside	Did not find that $\mu \neq \mu_0$: Fail to reject H_0	$\mu = \mu_0$ (assume)

UNIT 01 - SUMMARY

Visualization

VISUALIZATION

https://python-graph-gallery.com



UNIT 01 - SUMMARY

Let’s Practice…

STATS SUMMARY

- Open Lesson05-Review-Starter
- Data is in “datasets”

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET

COURSE

BEFORE NEXT CLASS

BEFORE NEXT CLASS

Start Working …

- Project: Unit Project 2
- Think about Final Project …

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS	~~What is Data Science~~	~~Lesson 1~~	
	~~Research Design and Pandas~~	~~Lesson 2~~	
	~~Statistics Fundamentals I~~	~~Lesson 3~~	
	~~Statistics Fundamentals II~~	~~Lesson 4~~	
	~~Flexible Class Session~~	~~Lesson 5~~	
UNIT 2: FOUNDATIONS OF DATA MODELING	Introduction to Regression	Lesson 6	Next Class
	Evaluating Model Fit	Lesson 7	
	Introduction to Classification	Lesson 8	
	Introduction to Logistic Regression	Lesson 9	
	Communicating Logistic Regression Results	Lesson 10	
	Flexible Class Session	Lesson 11	
UNIT 3: DATA SCIENCE IN THE REAL WORLD	Decision Trees and Random Forests	Lesson 12	
	Natural Language Processing	Lesson 13	
	Dimensionality Reduction	Lesson 14	
	Time Series Data I	Lesson 15	
	Time Series Data II	Lesson 16	
	Database Technologies	Lesson 17	
	Where to Go Next	Lesson 18	
	Flexible Class Session	Lesson 19	
	Final Project Presentations	Lesson 20	

LESSON

Q & A