

STATISTICS FUNDAMENTALS

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

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



Today's Class

LAST CLASS

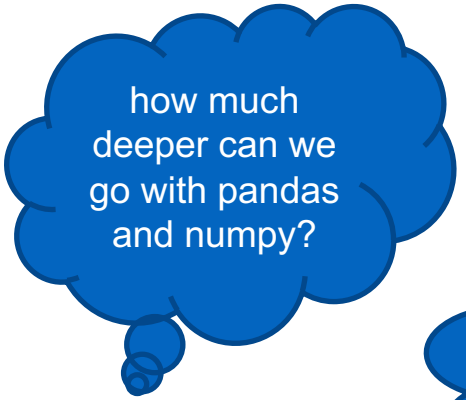
WHAT DID WE LEARN?

- ✓ Manage your development environment and files
- ✓ Define and Identify a problem and types of data
- ✓ Apply the data science workflow in the pandas context
- ✓ Create an Notebook to import, format, and clean using the Pandas

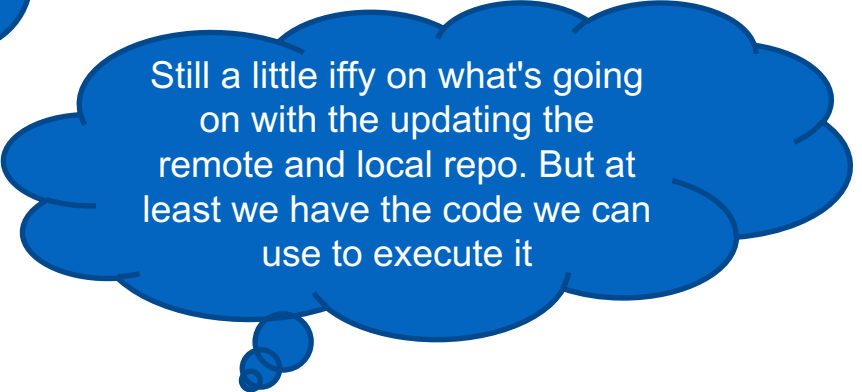
LAST CLASS

ANNOUNCEMENTS


- ❖ We need to talk. Reserve your 1:1 on doodle (Last Call)
- ❖ Happy Hour ... Are you interested?
- ❖ Fill your exit ticket!

A blue thought bubble with a small tail pointing towards the bottom left.

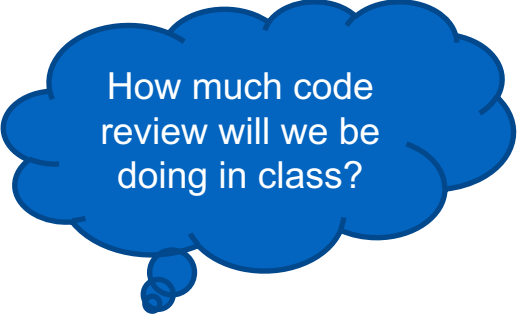
how much deeper can we go with pandas and numpy?

A large blue thought bubble with a small tail pointing towards the bottom center.

Still a little iffy on what's going on with the updating the remote and local repo. But at least we have the code we can use to execute it

A blue thought bubble with a small tail pointing towards the bottom left.

cant think of one

A blue thought bubble with a small tail pointing towards the bottom left.

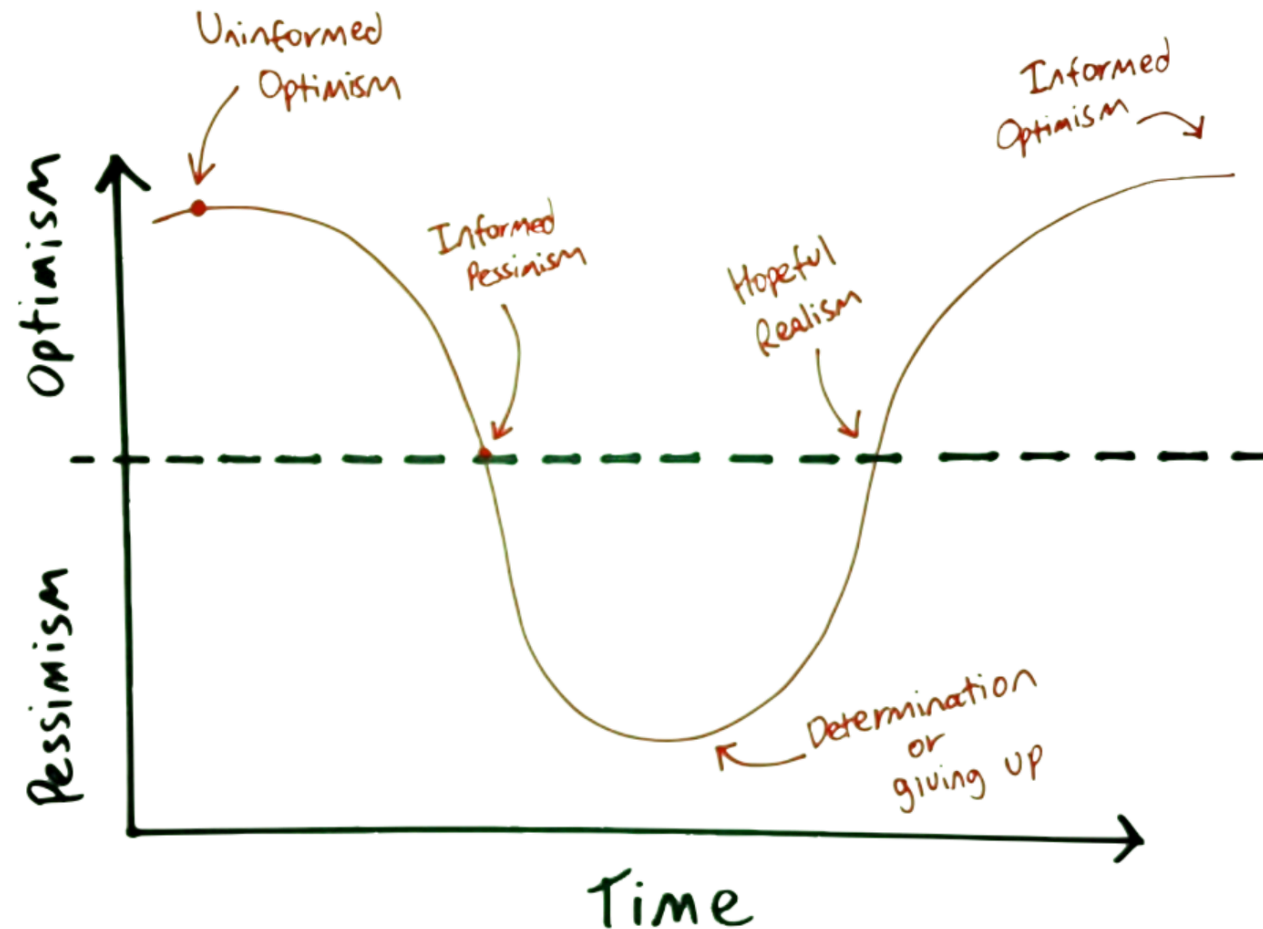
How much code review will we be doing in class?

A blue thought bubble with a small tail pointing towards the bottom left.

Others?

CLASS EXPECTATIONS

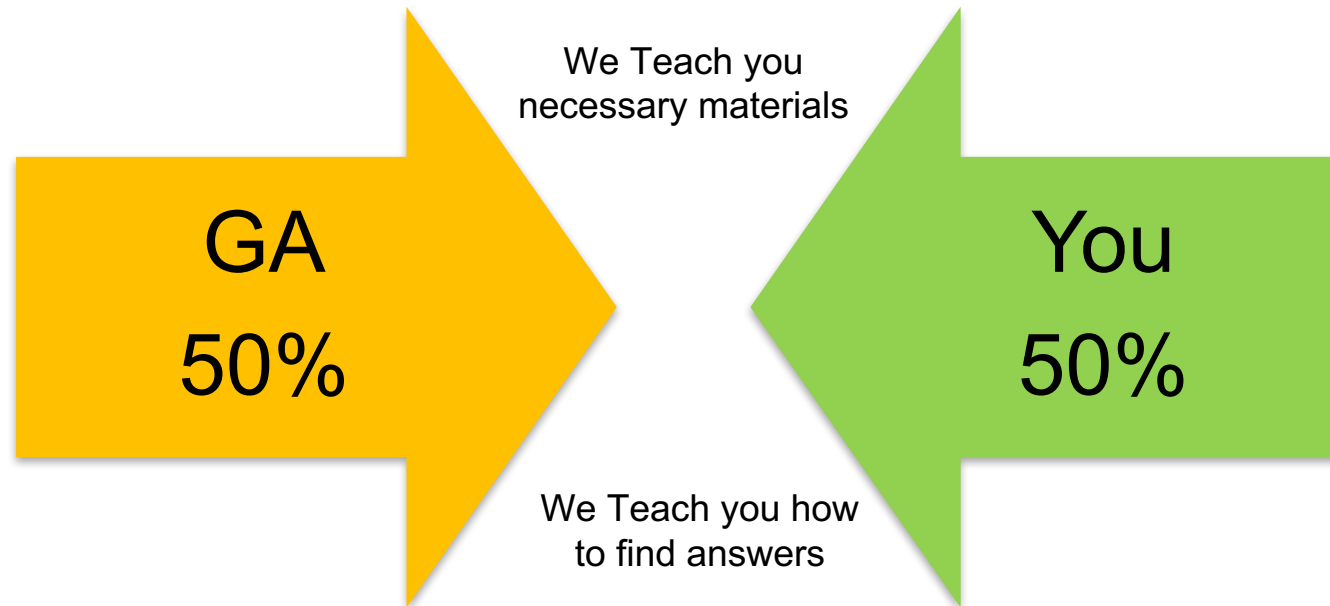
Where Are You Now?



CLASS EXPECTATIONS

Student Responsibilities

As a self directed program, you are a crucial part of the skill acquisition process.



How many hours per week are you spending at home preparing and practicing?

AFTER LAST CLASS

Do you know how to do the following?

- Manage your files using Git and GitHub?

```
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
```

Note

If you are asked to edit after committing you case use the following:

- `git config --global core.editor "vim"` (only one time)
- `ESC:q!` (to close the editor)

AFTER LAST CLASS

Do you know how to do the following?

- Open your Spyder/Jupyter?
- Do basics pandas like: `read_csv`, `describe()`, sorting?

AFTER LAST CLASS

Do you know how to do the following?

- Git
- Open your Spyder/Jupyter?
- Do basics pandas like: `read_csv`, `describe()`, sorting?



You got all 3 objectives?



Not all of them...

Let's form groups of 1's and 2's ...

STATISTICS FUNDAMENTALS

LEARNING OBJECTIVES

- Review basic pandas functions using lab2, a new dataset, and unit project
- Use NumPy and Pandas libraries to analyze datasets using basic summary statistics
- Create basic data visualizations to discern characteristics and trends in a dataset
- Identify a normal distribution within a dataset using summary statistics and visualization
- ID variable types and complete dummy coding by hand

REVIEW

PANDAS REVIEW

PANDAS QUIZ – Answer 10 questions in 10 minutes

1. Create this dataframe and call it df
2. Order rows by values of “a” low to high
3. Change df so the order of rows are sorted by values of “c” high to low
4. Rename column ”b” to “mpg”
5. Create df1 with only column “a” that you want to change its name to “model”
6. Create df2 where you append columns of df1 to df
7. Drop column c from df2 and then drop duplicates
8. Extract rows from df2 with mpg ≤ 3
9. Show first row and last row of df2
10. Show summary of statistics from df2

a	b	c
2	4	5
8	3	2
4	3	7

PANDAS REVIEW – LAB 2 QUESTIONS?

Lab 2 Solution

This is a quiz given in Roger Peng's [Coursera](#) class [Computing for Data Analysis](#).

```
In [1]: import pandas as pd
import os

data = pd.read_csv(os.path.join('..', '..', 'assets', 'dataset', 'ozone.csv'))
```

```
In [2]: print data.head()
```

	Ozone	Solar.R	Wind	Temp	Month	Day
0	41	190	7.4	67	5	1
1	36	118	8.0	72	5	2
2	12	149	12.6	74	5	3
3	18	313	11.5	62	5	4
4	NaN	NaN	14.3	56	5	5

Print the column names of the dataset to the screen, one column name per line.

```
In [3]: for x in data.columns.values:
        print x
```

Ozone
Solar.R
Wind
Temp
Month
Day

Extract the first 2 rows of the data frame and print them to the console. What does the output look like?

```
In [4]: tmp = data.ix[0:1] # or data.head(2)
print tmp.head()
```

	Ozone	Solar.R	Wind	Temp	Month	Day
0	41	190	7.4	67	5	1
1	36	118	8.0	72	5	2

Did you make your own
Pandas Cheat Sheet?

PANDAS REVIEW – LAB 2 QUESTIONS?

How many observations (i.e. rows) are in this data frame?

```
In [5]: print len(data)
```

153

Extract the last 2 rows of the data frame and print them to the console. What does the output look like?

```
In [6]: tmp = data.tail(2)
print tmp.head()
```

	Ozone	Solar.R	Wind	Temp	Month	Day
151	18	131	8.0	76	9	29
152	20	223	11.5	68	9	30

What is the value of Ozone in the 47th row?

```
In [7]: print data.ix[46:48,]
```

	Ozone	Solar.R	Wind	Temp	Month	Day
46	21	191	14.9	77	6	16
47	37	284	20.7	72	6	17
48	20	37	9.2	65	6	18

```
In [8]: print data.ix[47, 'Ozone']
```

37.0

How many missing values are in the Ozone column of this data frame?

```
In [9]: print data['Ozone'].isnull().sum()
print len(data) - len(data['Ozone'].dropna())
```

37
37

What is the mean of the Ozone column in this dataset? Exclude missing values (coded as NA) from this calculation.

```
In [10]: print data['Ozone'].mean()
```

42.1293103448

Did you make your own
Pandas Cheat Sheet?

PANDAS REVIEW – LAB 2 QUESTIONS?

Extract the subset of rows of the data frame where Ozone values are above 31 and Temp values are above 90. What is the mean of "Solar.R" in this subset?

```
In [11]: print data[(data.Ozone > 31) & (data.Temp > 90)].head()
```

	Ozone	Solar.R	Wind	Temp	Month	Day
68	97	267	6.3	92	7	8
69	97	272	5.7	92	7	9
119	76	203	9.7	97	8	28
120	118	225	2.3	94	8	29
121	84	237	6.3	96	8	30

```
In [12]: print data[(data.Ozone > 31) & (data.Temp > 90)][ 'Solar.R' ].mean()
```

212.8

What is the mean of "Temp" when "Month" is equal to 6?

```
In [13]: print data[ data.Month==6 ].Temp.mean()  
print data[ data.Month==6 ][ 'Temp' ].mean()
```

79.1
79.1

What was the maximum ozone value in the month of May (i.e. Month = 5)?

```
In [14]: print data[ data.Month==5 ].Ozone.max()
```

115.0

Did you make your own
Pandas Cheat Sheet?

PANDAS REVIEW – PRACTICE PROBLEM (5 min)

1. Import data “zillow” as a dataframe
2. Show the first 5 rows
3. Describe the data
4. Create a new dataframe with these columns: address, latitude, longitude
5. Plot a histogram of the longitude (note: in Jupyter use *%matplotlib inline*)

PANDAS REVIEW – PRACTICE PROBLEM

- Import data “zillow” as a dataframe

```
import pandas as pd
```

```
zillow_df = pd.read_csv("lessons/lesson-03/assets/dataset/zillow.csv")
```

- Show the first 5 rows

```
zillow_df.head()
```

- Describe the data

```
zillow_df.describe()
```

- Create a new dataframe with these columns: address, latitude, longitude

```
Df2 = zillow_df[["Address", "Latitude", "Longitude"]]
```

- Plot a histogram of the longitude

```
%matplotlib inline
```

```
Df2.Longitude.plot.hist()
```

PANDAS REVIEW – UNIT PROJECT 1 (DUE NEXT SESSION)

Exercise 1: Read and evaluate the following problem statement:

Determine which free-tier customers will covert to paying customers, using demographic data collected at signup (age, gender, location, and profession) and customer useage data (days since last log in, and activity score 1 = active user, 0= inactive user) based on Hooli data from Jan-Apr 2015.

1. What is the outcome?

Answer:

2. What are the predictors/covariates?

Answer:

3. What timeframe is this data relevent for?

Answer:

4. What is the hypothesis?

Answer:

PANDAS REVIEW – UNIT PROJECT 1 (DUE NEXT SESSION)

Exercise 2: Let's get started with our dataset (use admissions.csv)

1. Create a data dictionary

Answer:

Variable	Description	Type of Variable
Var 1	0 = not thing 1 = thing	categorical
Var 2	thing in unit X	continuous

We would like to explore the association between X and Y

2. What is the outcome?

Answer:

3. What are the predictors/covariates?

Answer:

4. What timeframe is this data relevant for?

Answer:

5. What is the hypothesis?

Answer:

6. Using the above information, write a well-formed problem statement.

Answer:

PANDAS REVIEW – UNIT PROJECT 1 (DUE NEXT SESSION)

Exercise 3: Exploratory Analysis Plan (Materials will be covered on Tuesday's class)

Using the lab from a class as a guide, create an exploratory analysis plan.

1. What are the goals of the exploratory analysis?

Answer:

2a. What are the assumptions of the distribution of data?

Answer:

2b. How will determine the distribution of your data?

Answer:

3a. How might outliers impact your analysis?

Answer:

3b. How will you test for outliers?

Answer:

4a. What is colinearity?

Answer:

4b. How will you test for colinearity?

Answer:

5. What is your exploratory analysis plan?

Using the above information, write an exploratory analysis plan that would allow you or a colleague to reproduce your analysis 1 year from now.

Answer:

STATISTICS FUNDAMENTALS

LEARNING OBJECTIVES

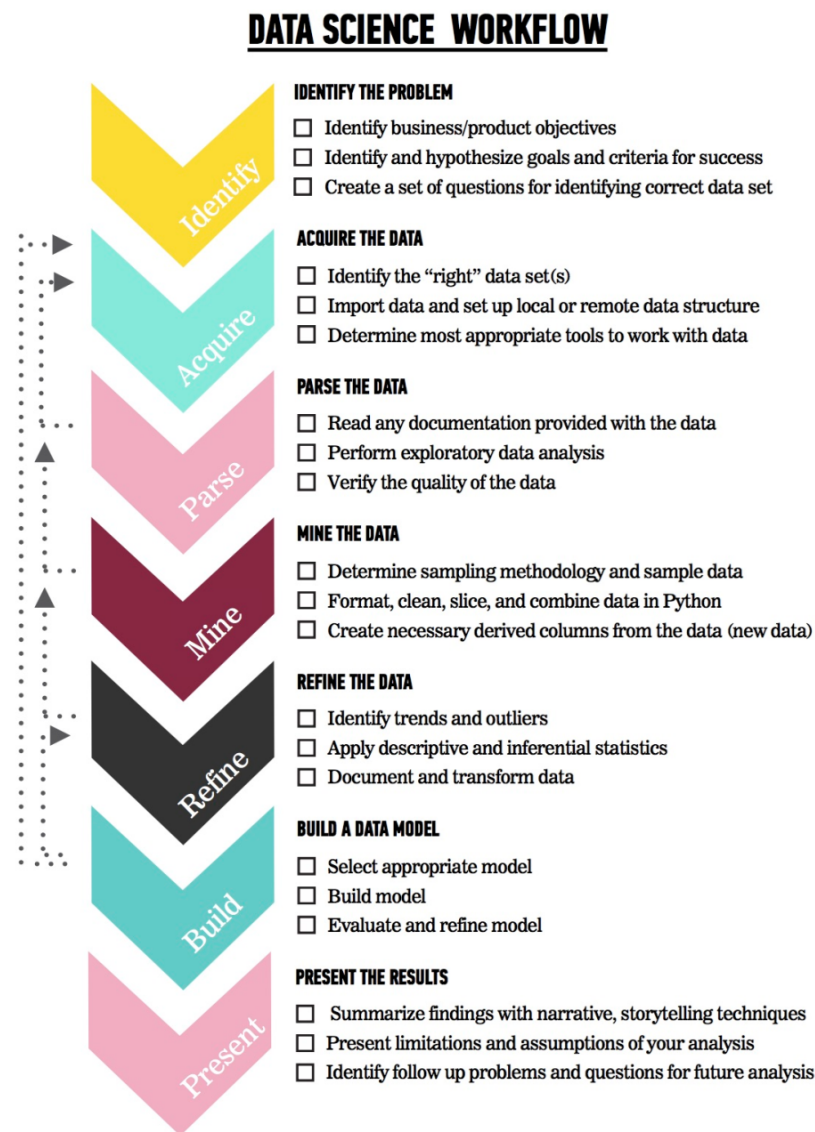
- ~~Review basic pandas functions using lab2, a new dataset, and unit project~~
- Use NumPy and Pandas libraries to analyze datasets using basic summary statistics
- Create basic data visualizations to discern characteristics and trends in a dataset
- Identify a normal distribution within a dataset using summary statistics and visualization
- ID variable types and complete dummy coding by hand

STATISTICS FUNDAMENTALS

LET'S REVIEW THE DATA SCIENCE WORKFLOW

The steps:

1. Identify the problem
2. Acquire the data
3. Parse the data
4. Mine the data
5. Refine the data
6. Build a data model
7. Present the results



TODAY

- We're going to elaborate more on step 3: Parsing the Data
- We'll begin to talk about the fundamentals of Statistics

INTRODUCTION

LAYING THE GROUND WORK

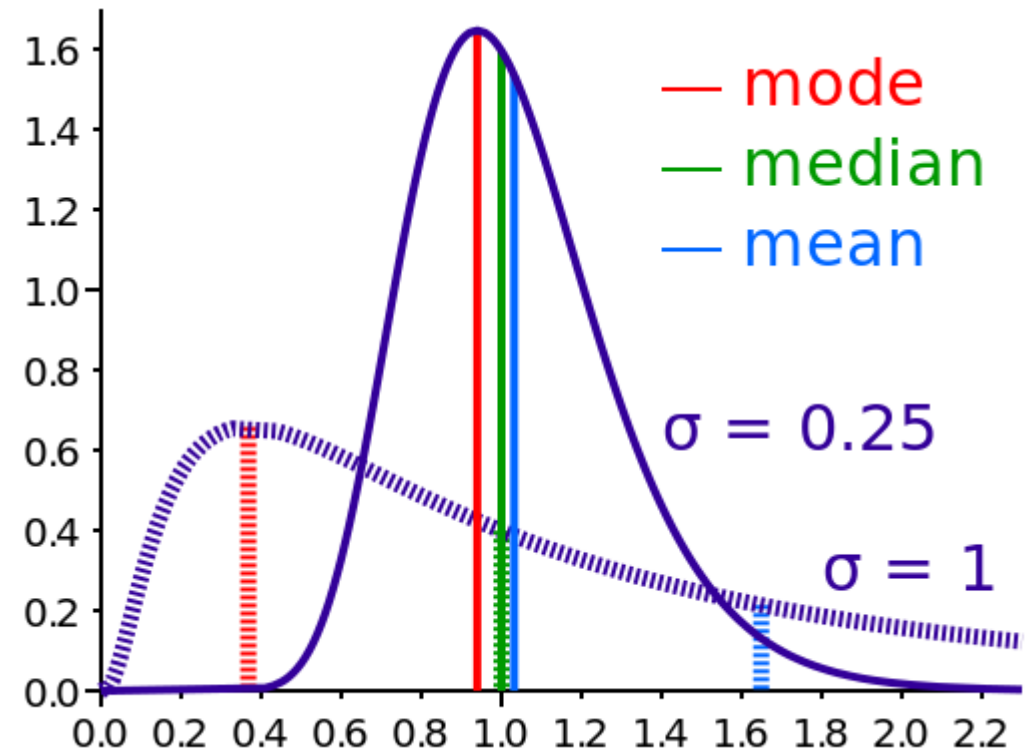
WE'RE GOING TO COVER SEVERAL TOPICS

- Mean
- Median
- Mode
- Max
- Min
- Quartile
- Interquartile Range
- Variance
- Standard Deviation
- Correlation

MEAN

- The mean of a set of values is the sum of the values divided by the number of values. It is also called the average.

$$\bar{X} = \frac{\sum X}{N}$$



MEAN EXAMPLE

- Find the mean of 19, 13, 15, 25, and 18.

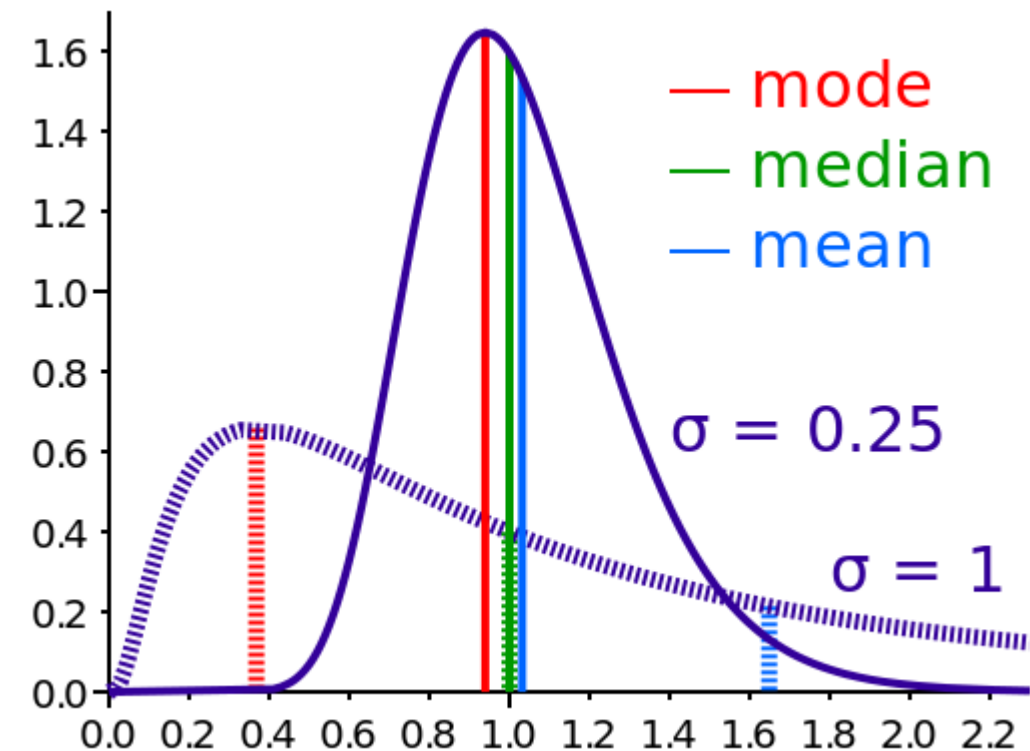
MEAN EXAMPLE

- Find the mean of 19, 13, 15, 25, and 18.

$$\frac{19 + 13 + 15 + 25 + 18}{5} = \frac{90}{5} = 18$$

MEDIAN

- The median refers to the midpoint in a series of numbers.
- To find the median
 - Arrange the numbers in order smallest to largest.
 - If there is an odd number of values, the middle value is the median.
 - If there is an even number of values, the average of the middle two values is the median.



MEDIAN EXAMPLE

- Find the median of 19, 29, 36, 15, and 20.

MEDIAN EXAMPLE

- Find the median of 19, 29, 36, 15, and 20.

Ordered Values:

15, 19, 20, 29, 36

20 is the median

MEDIAN EXAMPLE

- Find the median of 67, 28, 92, 37, 81, 75.

MEDIAN EXAMPLE

- Find the median of 67, 28, 92, 37, 81, 75.

Ordered Values:

28, 37, 67, 75, 81, 92

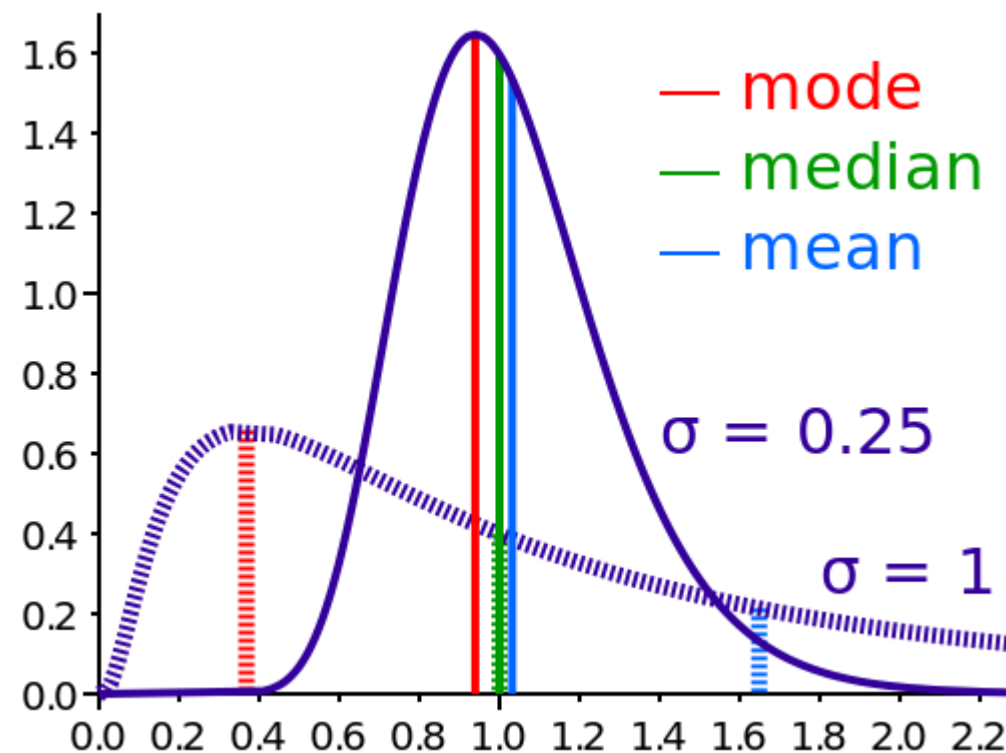
67 and 75 are the middle values.

$$\frac{67 + 75}{2} = \frac{142}{2} = 71$$

71 is the median.

MODE

- The mode of a set of values is the value that occurs most often.
- A set of values may have more than one mode or no mode.

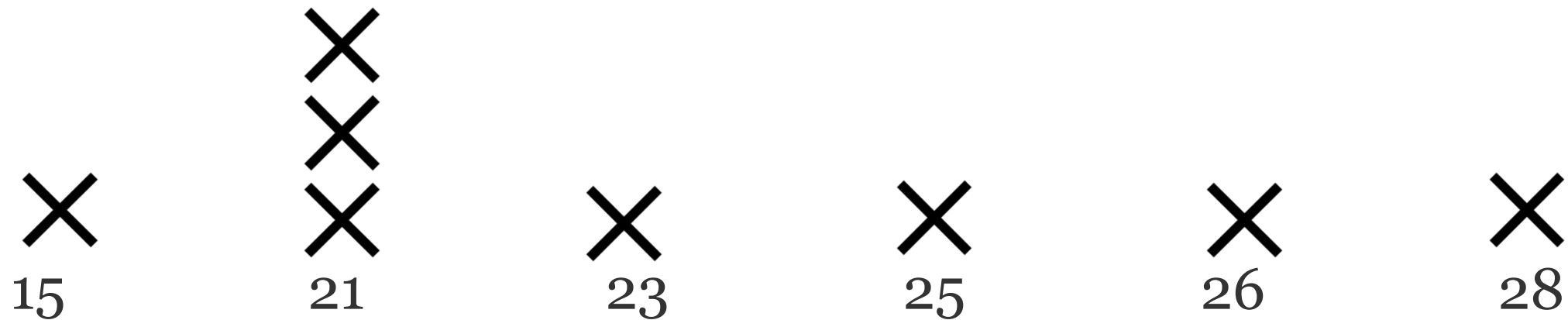


MODE EXAMPLE

- Find the mode of 15, 21, 26, 25, 21, 23, 28, and 21.

MODE EXAMPLE

- Find the mode of 15, 21, 26, 25, 21, 23, 28, and 21.



21 is the mode because it occurs most frequently

MODE EXAMPLE

- Find the mode of 12, 15, 18, 26, 15, 9, 12, and 27.

MODE EXAMPLE

- Find the mode of 12, 15, 18, 26, 15, 9, 12, and 27.



12 and 15 are the modes since the both occur twice.

MODE EXAMPLE

- Find the mode of 4, 8, 15, 21, and 23.

MODE EXAMPLE

- Find the mode of 4, 8, 15, 21, and 23.

✕
4

✕
8

✕
15

✕
21

✕
23

There is no mode since all values occur the same number of times.

ACTIVITY: KNOWLEDGE CHECK



EXERCISE

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. For the following groups of numbers, calculate the mean, median and mode by hand. Also determine the min and max.
 - a. 18, 24, 17, 21, 24, 16, 29, 18
 - b. 75, 87, 49, 68, 75, 84, 98, 92
 - c. 55, 47, 38, 66, 56, 64, 44, 39

DELIVERABLE

Answers to the above questions

ACTIVITY: KNOWLEDGE CHECK

ANSWERS



List	Mean	Median	Mode	Max	Min
18, 24, 17, 21, 24, 16, 29, 18	20.875	19.5	18 & 24	29	16
75, 87, 49, 68, 75, 84, 98, 92	78.5	79.5	75	98	49
55, 47, 38, 66, 56, 64, 44, 39	51.125	51	None	66	38

CODEALONG

SUMMARY STATISTICS IN PANDAS

CODEALONG: SUMMARY STATISTICS IN PANDAS

- Open the starter-code notebook located in lessons/lesson-03/code/starter-code of the class repo.

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

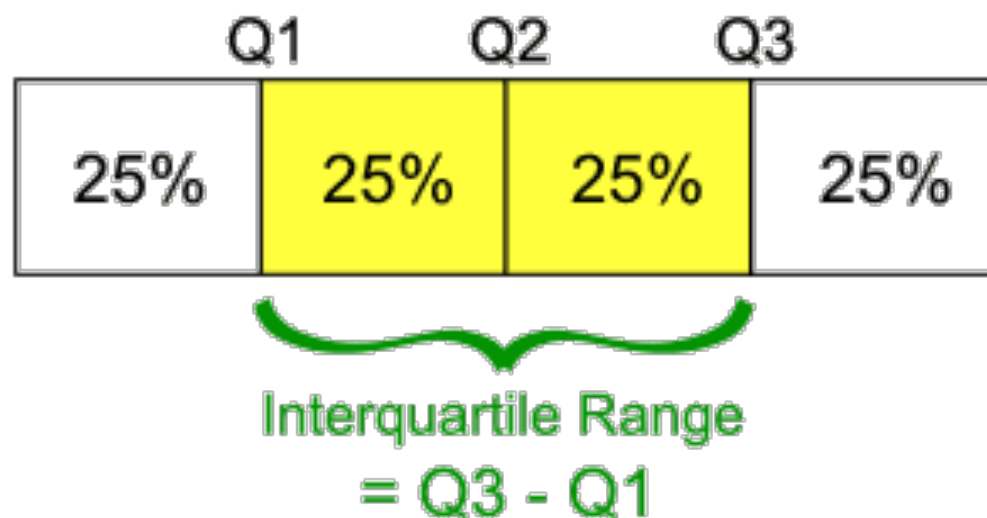
`.median()` - Compute median value

`.mode()` - Compute mode value

`.count()` - Count the number of observations

QUARTILES AND INTERQUARTILE RANGE

- Quartiles divide a rank-ordered data set into four equal parts.
- The values that divide each part are called first, second, and third quartiles, denoted $Q1$, $Q2$, and $Q3$, respectively.
- The interquartile range (IQR) is $Q3 - Q1$, a measure of variability.



Complete on your notebook Part 1

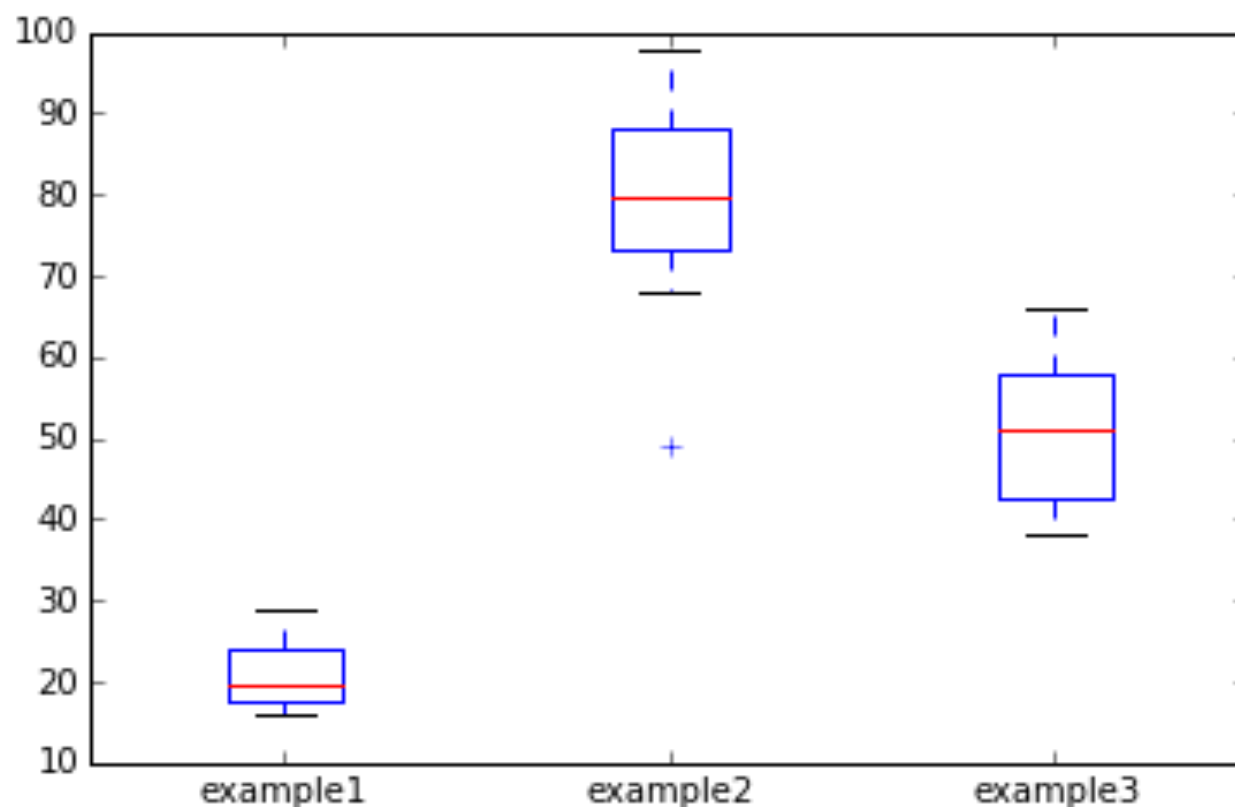
STATISTICS FUNDAMENTALS

LEARNING OBJECTIVES

- ~~▶ Review basic pandas functions using lab2, a new dataset, and unit project~~
- ~~▶ Use NumPy and Pandas libraries to analyze datasets using basic summary statistics~~
- ▶ Create basic data visualizations to discern characteristics and trends in a dataset
- ▶ Identify a normal distribution within a dataset using summary statistics and visualization
- ▶ ID variable types and complete dummy coding by hand

CODEALONG PART 2: BOX PLOT

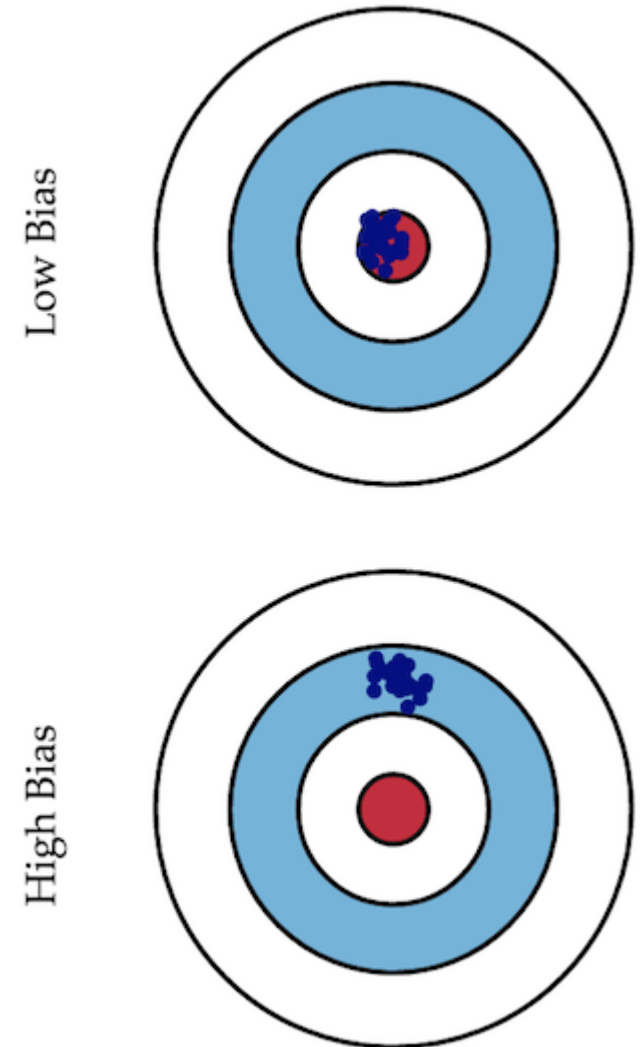
- Box plots give a nice visual of min, max, mean, median, and the quartile and interquartile range.



Complete on your notebook Part 2

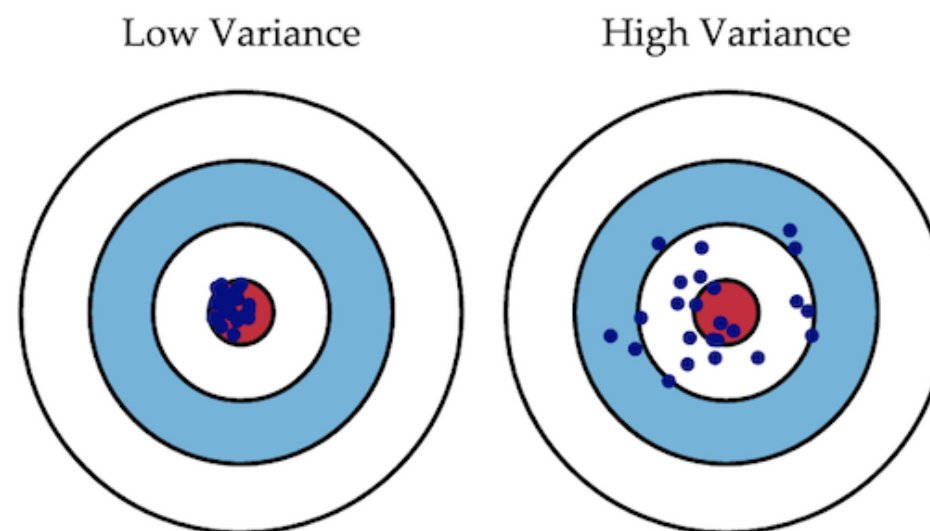
BIAS VS. VARIANCE

- Error due to **bias** is calculated at the difference between the *expected prediction* of our model and the *correct value* we are trying to predict.
- Imagine creating multiple models on various datasets. **Bias** measures *how far off in general* models' predictions are from the correct value.

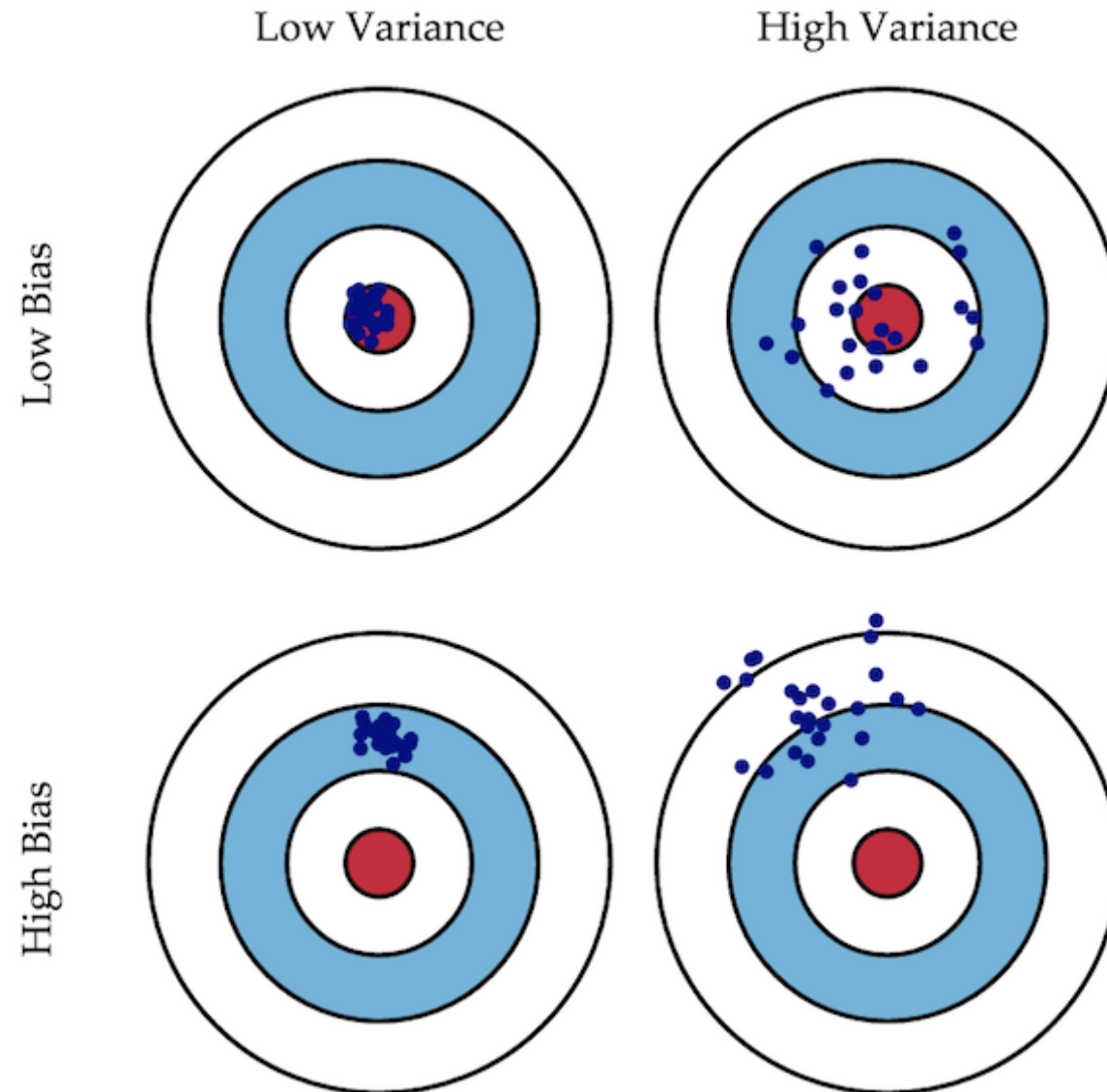


BIAS VS. VARIANCE

- Error due to **variance** is taken as the variability of a model prediction for a given point.
- Imagine creating multiple models on various datasets. The **variance** is *how much the predictions for a given point vary* between different realizations of the model.



BIAS VS. VARIANCE



STANDARD DEVIATION

- Standard deviation (SD, σ for population, s for sample) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.
- **Standard deviation** is the square root of **variance**.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Let's do it manually first

CODEALONG PART 3: STANDARD DEVIATION & VARIANCE

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

Methods include:

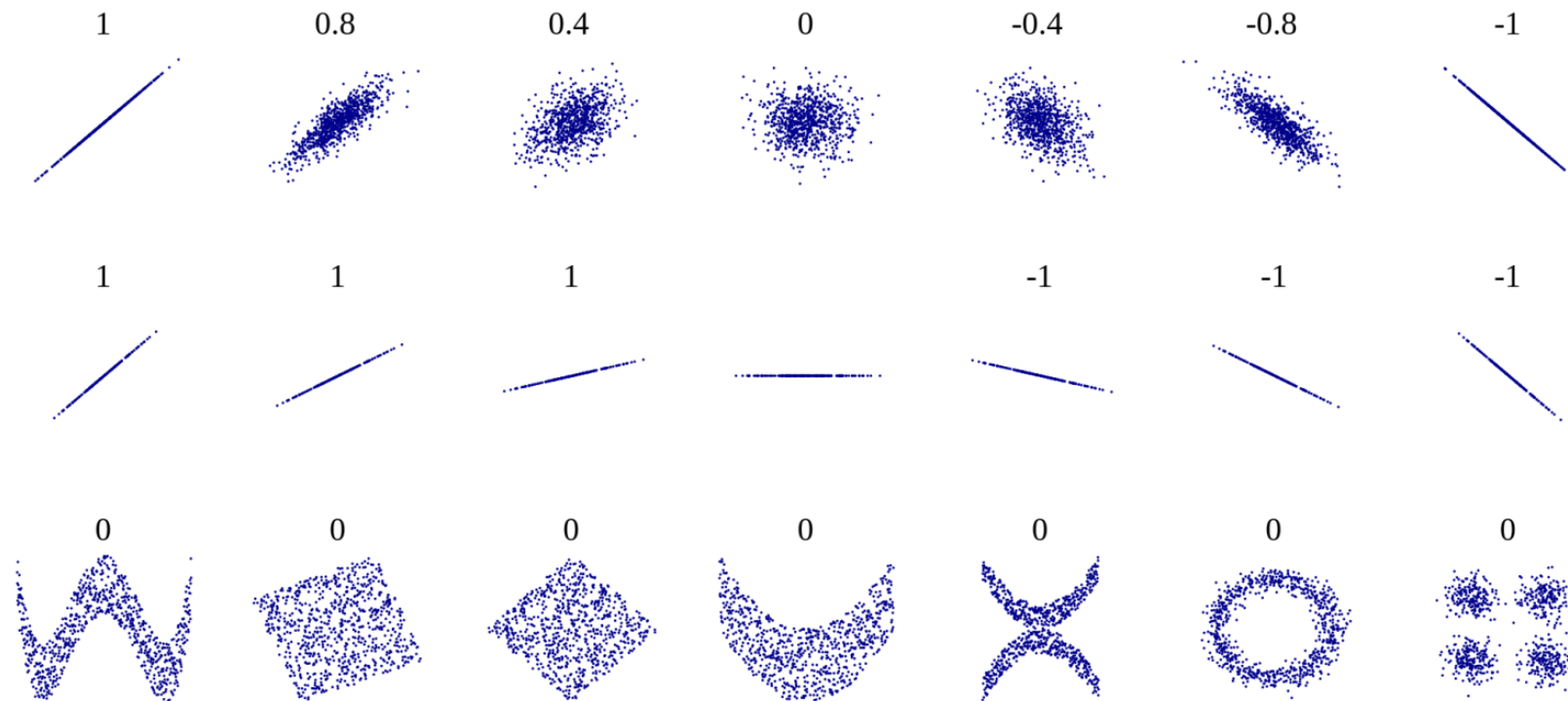
`.std()` - Compute Standard Deviation

`.var()` - Compute variance

`.describe()` - short cut that prints out count, mean, std, min, quartiles, max

CORRELATION

- The correlation measures the extent of interdependence of variable quantities.
- Example correlation values



CONTEXT

- For most projects, descriptive stats will come first. These help you get to know your dataset better.
- Sometimes, descriptive stats may be all you need to answer your question.

Complete on your notebook Part 4

STATISTICS FUNDAMENTALS

LEARNING OBJECTIVES

- ~~▶ Review basic pandas functions using lab2, a new dataset, and unit project~~
- ~~▶ Use NumPy and Pandas libraries to analyze datasets using basic summary statistics~~
- ~~▶ Create basic data visualizations to discern characteristics and trends in a dataset~~
- ▶ Identify a normal distribution within a dataset using summary statistics and visualization
- ▶ ID variable types and complete dummy coding by hand

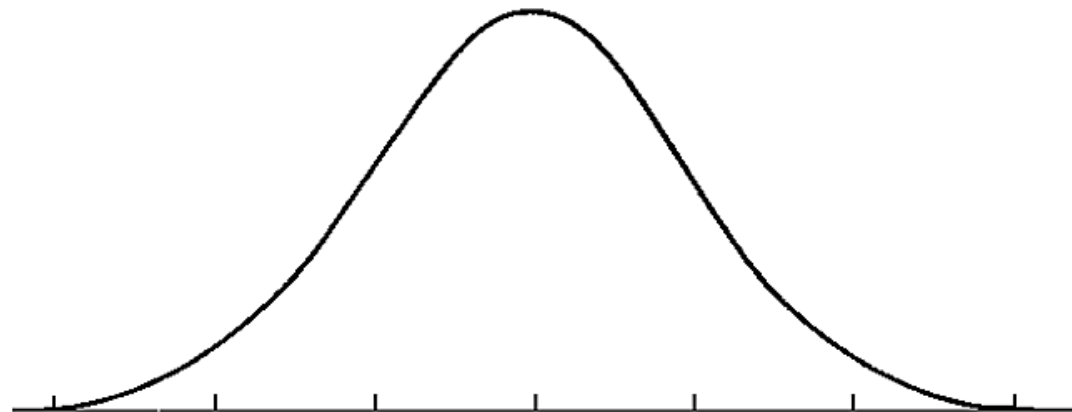
10 min
Break

INTRODUCTION

IS THIS NORMAL?

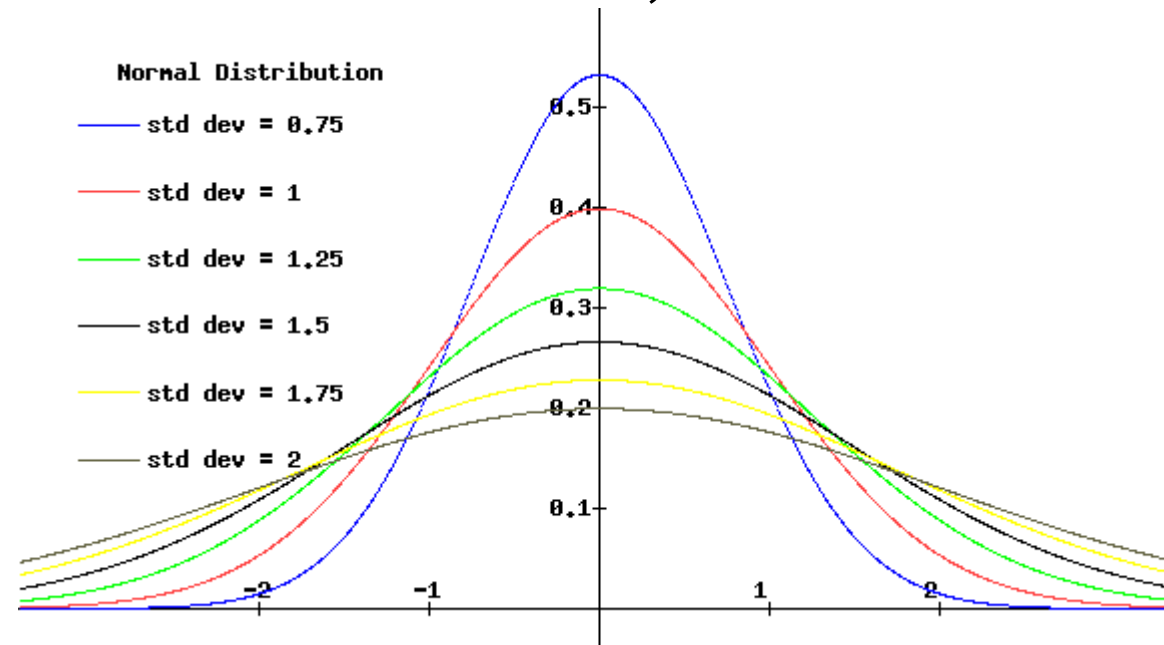
THE NORMAL DISTRIBUTION

- A normal distribution is often a key assumption to many models.
- The normal distribution depends upon the *mean* and the *standard deviation*.
- The *mean* determines the center of the distribution. The *standard deviation* determines the height and width of the distribution.



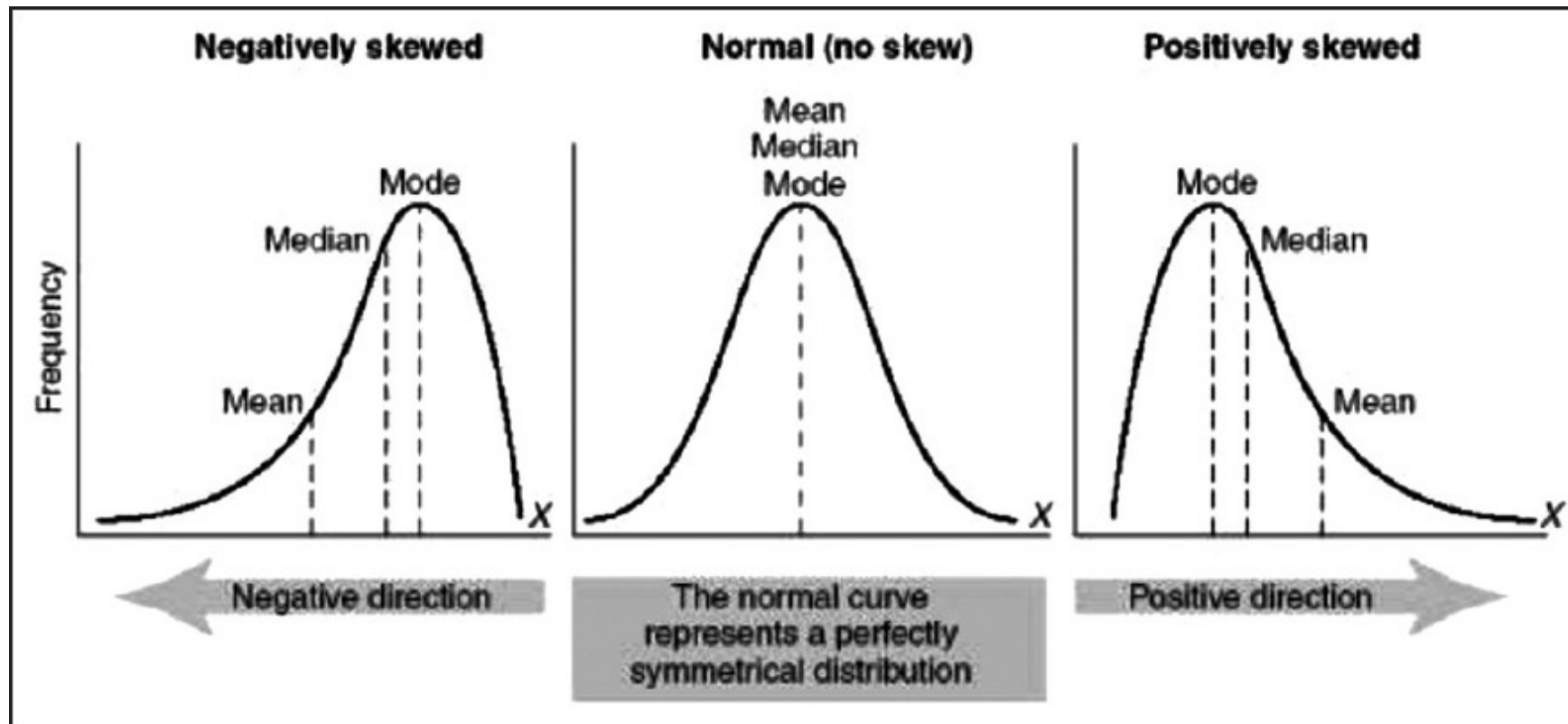
THE NORMAL DISTRIBUTION

- Normal distributions are symmetric, bell-shaped curves.
- When the standard deviation is large, the curve is short and wide.
- When the standard deviation is small, the curve is tall and narrow.



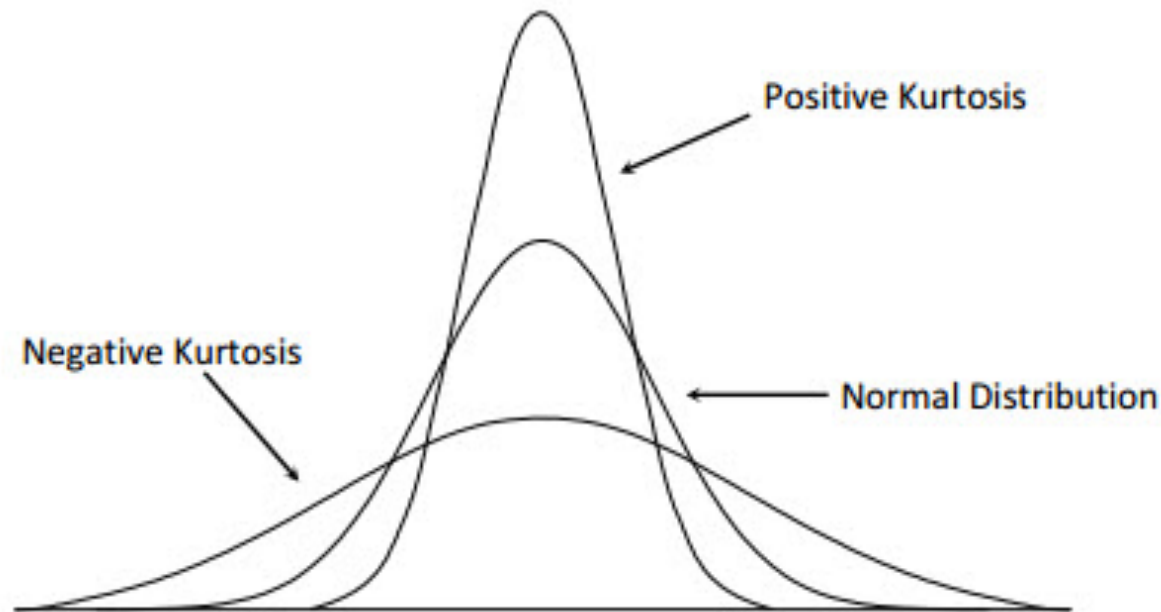
SKEWNESS

- 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.



DETERMINING THE DISTRIBUTION OF YOUR DATA

- Follow along as we walk through this in an iPython Notebook.
- Open Lesson-3-demo

Let's walk through Part 1 of lesson-3-demo together

STATISTICS FUNDAMENTALS

LEARNING OBJECTIVES

- ~~▶ Review basic pandas functions using lab2, a new dataset, and unit project~~
- ~~▶ Use NumPy and Pandas libraries to analyze datasets using basic summary statistics~~
- ~~▶ Create basic data visualizations to discern characteristics and trends in a dataset~~
- ~~▶ Identify a normal distribution within a dataset using summary statistics and visualization~~
- ▶ ID variable types and complete dummy coding by hand

INTRODUCTION

VARIABLE TYPES

VARIABLE TYPES

- Numeric variables can take on a large range of non-predetermined, quantitative values. These are things such as height, income, etc.
- Categorical variables can take on a specific set of variables. These are things such as race, gender, paint colors, movie titles, etc.

Let's walk through Part 2 of lesson-3-demo together

DEMO

CLASSES

CLASS/DUMMY VARIABLES

- Let's say we have the categorical variable `area`, which takes on one of the following values: `rural`, `suburban`, and `urban`.
- We need to represent these numerically for a model. So how do we code them?

CLASS/DUMMY VARIABLES

- How about 0=rural, 1=suburban, and 2=urban?

CLASS/DUMMY VARIABLES

- But this implies an ordered relationship - is urban twice suburban?
That doesn't make sense.
- However, we can represent this information by converting the one area variable into two new variables:

area_urban and area_suburban.

CLASS/DUMMY VARIABLES

- We'll draw out how categorical variables can be represented without implying order.
- First, let's choose a reference category. This will be our “base” category.
- It's often good to choose the category with the largest sample size and a criteria that will help model interpretation. If we are testing for a disease, the reference category would be people without the disease.

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`.

CLASS/DUMMY VARIABLES

- Why do we need only two dummy variables?

rural	urban	suburban
-------	-------	----------

- We can derive all of the possible values from these two. If an area isn't urban or suburban, we know it must be rural.
- In general, if you have a categorical feature with k categories, you need to create $k-1$ dummy variable to represent all of the information.

CLASS/DUMMY VARIABLES

- Let's see our dummy variables.

	area_urban	area_suburban
rural	0	0
suburban	0	1
urban	1	0

- As mentioned before, if we know $\text{area_urban}=0$ and $\text{area_suburban}=0$, then the area must be rural.

CLASS/DUMMY VARIABLES

- We can do this for a gender variable with two categories: male and female.
- How many dummy variables need to be created?

CLASS/DUMMY VARIABLES

▸ # of categories - 1 = 2 - 1 = 1

CLASS/DUMMY VARIABLES

- We will make `female` our reference category. Thus, `female=0` and `male=1`.

	gender_male
female	0
male	1

- This can be done in Pandas with the `get_dummies` method.

INDEPENDENT PRACTICE

DUMMY COLORS

ACTIVITY: DUMMY COLORS



EXERCISE

DIRECTIONS

It's important to understand the concept before we use the Pandas function `get_dummies` to create dummy variables. So today, we'll create our dummy variables by hand.

1. Draw a table like the one on the white board.
2. Create dummy variables for the variable “colors” that has 6 categories: blue, red, green, purple, grey, and brown. Use grey as the reference.

DELIVERABLE

Dummy variables table for colors

STATISTICS FUNDAMENTALS

LEARNING OBJECTIVES

- ~~▶ Review basic pandas functions using lab2, a new dataset, and unit project~~
- ~~▶ Use NumPy and Pandas libraries to analyze datasets using basic summary statistics~~
- ~~▶ Create basic data visualizations to discern characteristics and trends in a dataset~~
- ~~▶ Identify a normal distribution within a dataset using summary statistics and visualization~~
- ~~▶ ID variable types and complete dummy coding by hand~~

CONCLUSION

TOPIC REVIEW

REVIEW

- We talked about several different types of summary statistics, what are they?
- We covered several different types of visualizations; which ones?
- We talked about the normal distribution; how do we determine your data's distribution?
- Any other questions?

COURSE

**BEFORE NEXT
CLASS**

LESSON

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET

BEFORE NEXT CLASS

Start Working ...

- Project: Unit Project 2
- Think about Final Project ...
- Any requests for schedule change?

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
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



Today's Class

LESSON

Q & A