GENERAL ASSEMBLY

STATISTICS FUNDAMENTALS I

Angelo Klin

Katra Analytics

LEARNING OBJECTIVES

- Use NumPy and Pandas libraries to analyse datasets using basic summary statistics:
  - mean, median, mode, max, min, quartile, inter-quartile range, variance, standard deviation and correlation
- Create data visualisations, including line graphs, box plots and histograms, to discern characteristics and trends in a dataset
- Identify a Normal Distribution within a dataset using summary statistics and visualisation
- Identify variable types and complete dummy coding by hand

DATA SCIENCE

PRE-WORK

PRE-WORK REVIEW

- Create and open an Jupyter Notebook
- Complete the Python pre-work

OPENING

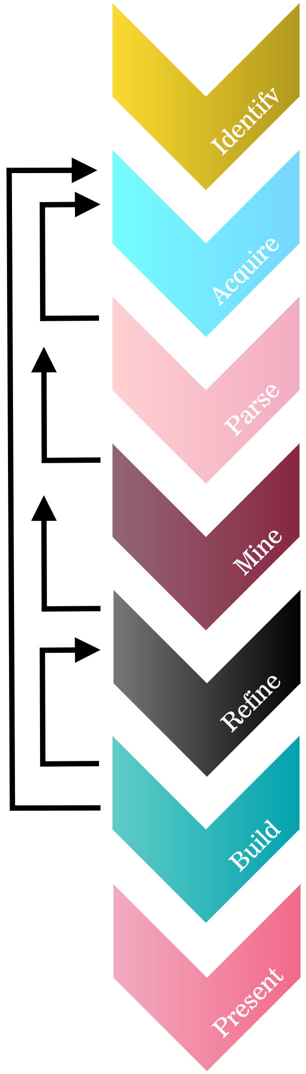
STATISTICS FUNDAMENTALS I

OPENING

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 are going to begin to talk about step 3
  - Parsing the Data
- We will begin to talk about the Fundamentals of Statistics

INTRODUCTION

LAYING THE GROUND WORK

TERMINOLOGY

Population	Sample
The Whole Every member of a group of interest	A Representative Part Some members of a group of interest
Has Parameters which describe the Population and do not change as long as the Population does not change	Has Statistics which approximate the Population Parameters
Use Greek letters or uppercase in equations	Use Roman letters or lowercase in equations

TERMINOLOGY

Measurement	Parameter (Use Greek letters or uppercase)	Statistic (Use Roman letters or lowercase)
Proportion	P	p
Data Elements	X	x
Number of Elements	N	n
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Correlation Coefficient	ρ	r

WE ARE GOING TO COVER SEVERAL TOPICS

- Measures of Central Tendency
  - Mean
  - Median
  - Mode
- Max
- Min
- Correlation
- Measures of Dispersion
  - Quartile
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

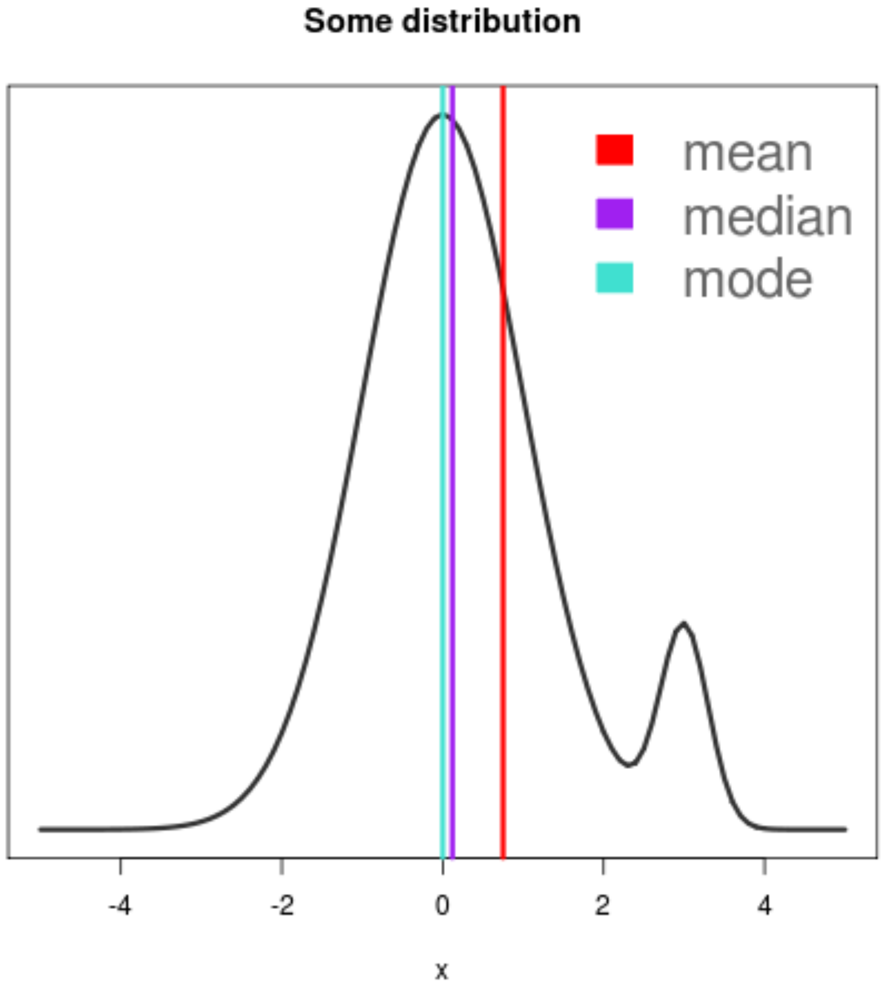
MEAN

- 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
- **Population Mean** represents the actual mean of the whole population

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

- **Sample Mean** is the arithmetic mean of random sample values drawn from the population

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$



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
  - The “balancing point” in a distribution
  - 50 percent of observations above and below

- To find the Median
  - Arrange the numbers **in order** smallest to largest
  - If there is an **odd** number of values
    - The Median is the middle value
  - If there is an **even** number of values
    - The Median is the average of the two middle values

MEDIAN EXAMPLE 1

- Find the Median of 19, 29, 36, 15 and 20

MEDIAN EXAMPLE 1

- Find the Median of 19, 29, 36, 15 and 20
  - Ordered Values
    - 15, 19, 20 , 29, 36

MEDIAN EXAMPLE 1

- Find the Median of 19, 29, 36, 15 and 20
  - Ordered Values
    - 15, 19, 20 , 29, 36
  - Odd number of values

MEDIAN EXAMPLE 1

- Find the Median of 19, 29, 36, 15 and 20
  - Ordered Values
    - 15, 19, 20 , 29, 36
  - Odd number of values
    - The Median is 20

MEDIAN EXAMPLE 2

- Find the Median of 67, 28, 92, 37, 81 and 75

MEDIAN EXAMPLE 2

- Find the Median of 67, 28, 92, 37, 81 and 75
  - Ordered Values
    - 28, 37, 67, 75, 81, 92

MEDIAN EXAMPLE 2

- Find the Median of 67, 28, 92, 37, 81 and 75

  - Ordered Values
    - 28, 37, 67, 75, 81, 92

  - Even number of values

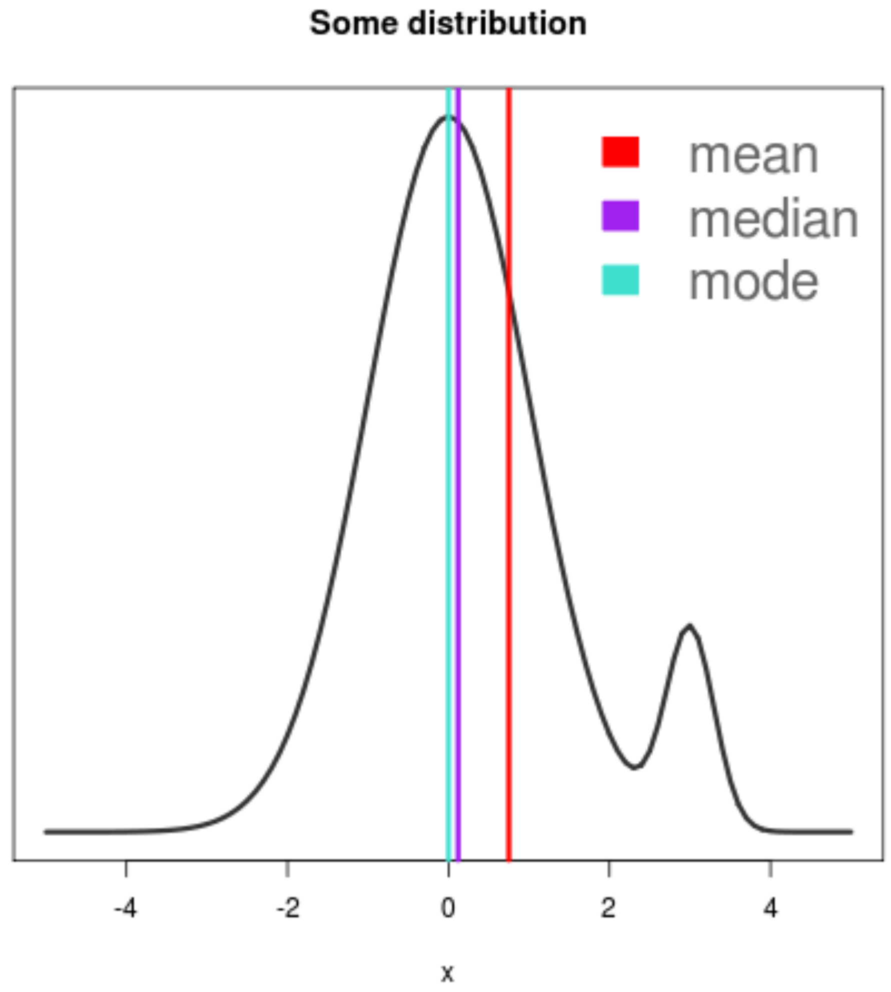
MEDIAN EXAMPLE 2

- Find the Median of 67, 28, 92, 37, 81 and 75
  - Ordered Values
    - 28, 37, 67, 75, 81, 92
  - Even number of values
    - The Median is (67 + 75) / 2
    - The Median is 71

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 1

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

MODE EXAMPLE 1

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

Number	Frequency
15	1
21	3
23	1
25	1
26	1
28	1

- The Mode is 21 because it occurs more frequently

MODE EXAMPLE 2

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

MODE EXAMPLE 2

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

Number	Frequency
9	1
12	2
15	2
18	1
26	1
27	1

- Both 12 and 15 are the Modes since they both occur twice

MODE EXAMPLE 3

- Find the Mode of 4, 8, 15, 21 and 23

MODE EXAMPLE 3

- Find the Mode of 4, 8, 15, 21 and 23

Number	Frequency
4	1
8	1
15	1
21	1
23	1

- There is No Mode since all values occur the same number of times

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

DIRECTIONS (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

CODE ALONG

SUMMARY STATISTICS IN PANDAS

CODE ALONG

- Open the starter code notebook located at
  - ~/lessons/lesson-03/code/starter/starter-3.ipynb

- Ask your classmates and instructor for help if you have problems!

CODE ALONG PART 1: BASIC STATISTICS

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

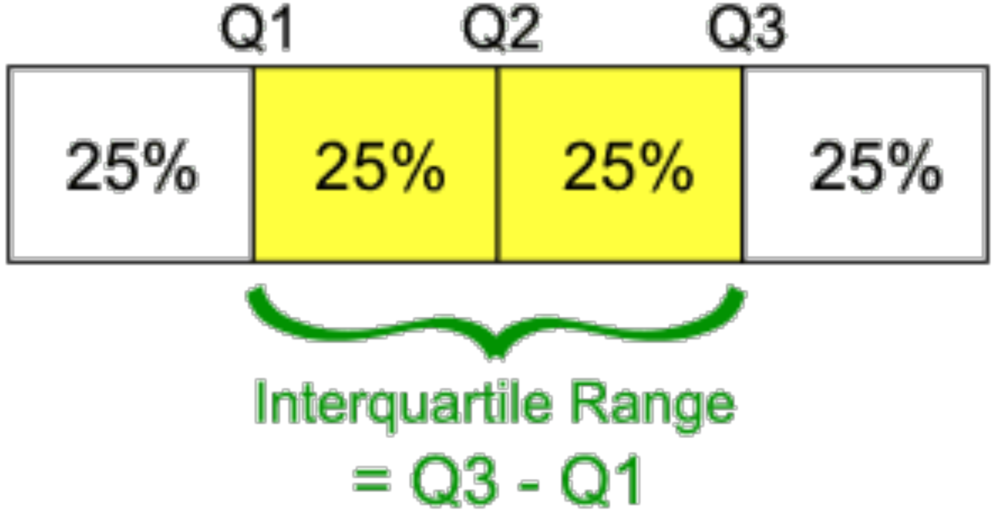
- Methods available include
  .min() - Compute minimum value
  .max() - Compute maximum value
  .mean() - Compute mean value
  .median() - Compute median value
  .mode() - Compute mode value
  .count() - Count the number of observations

QUARTILES AND INTER-QUARTILE RANGE

- Quartiles divide a rank-ordered dataset into four equal parts

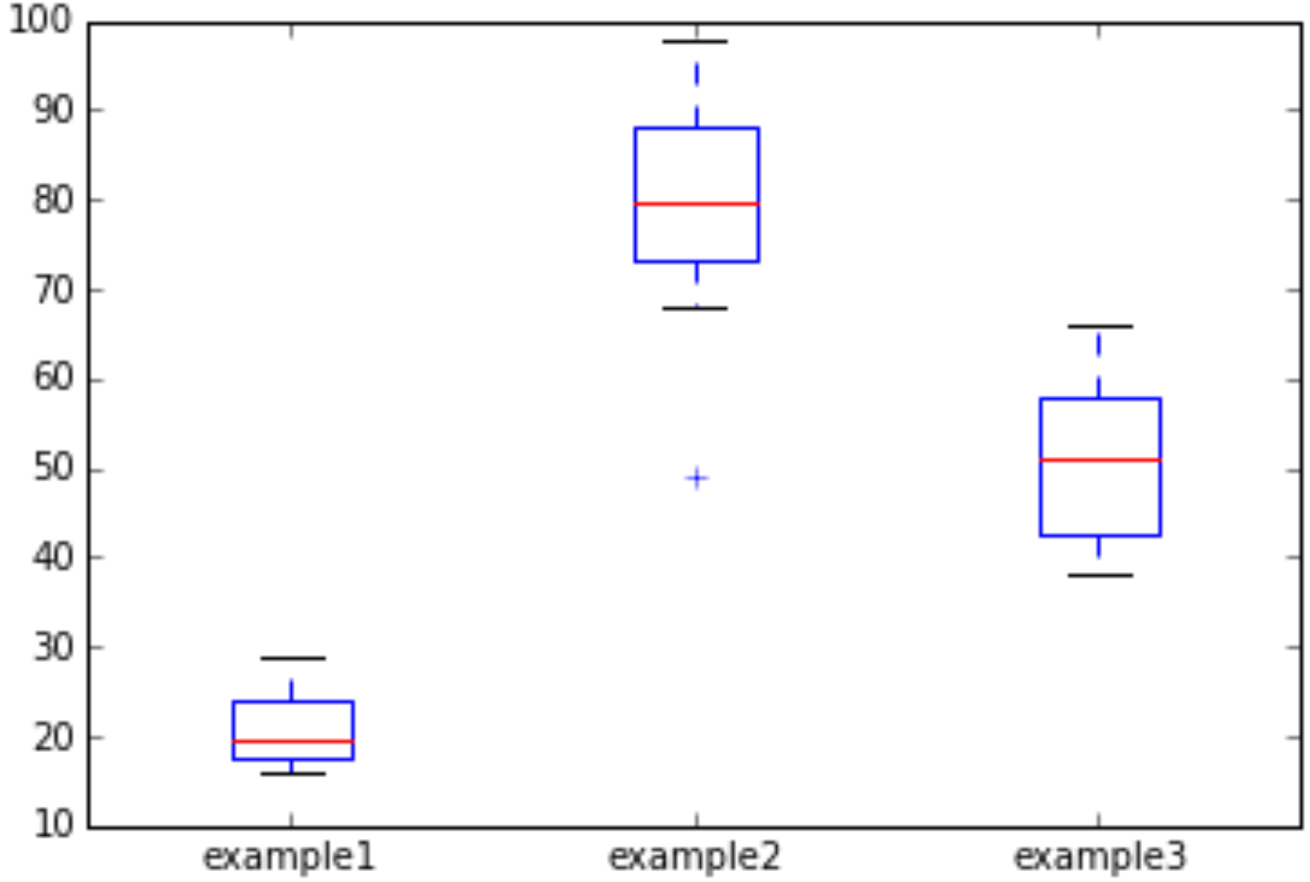
- The values that divide each part are called first, second, and third quartiles, denoted Q1, Q2 and Q3, respectively

- The interquartile range is a measure of variability; IQR = Q3 - Q1



CODE ALONG PART 2: BOX PLOT

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

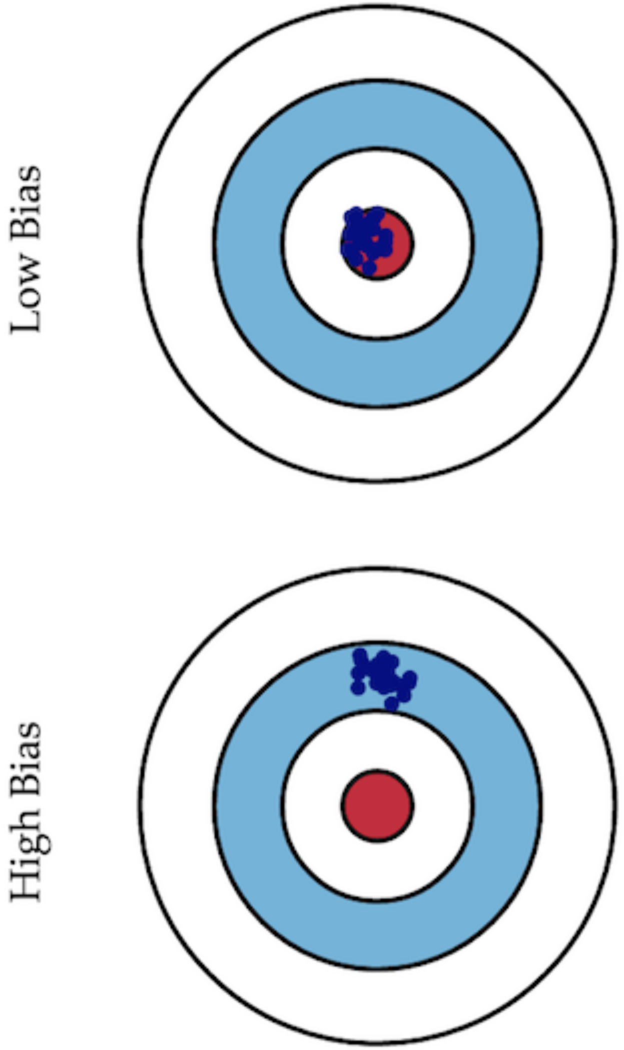


BIAS VS. VARIANCE

- Bias measures how far off the predictions are from the correct value

- Error due to Bias is calculated as 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

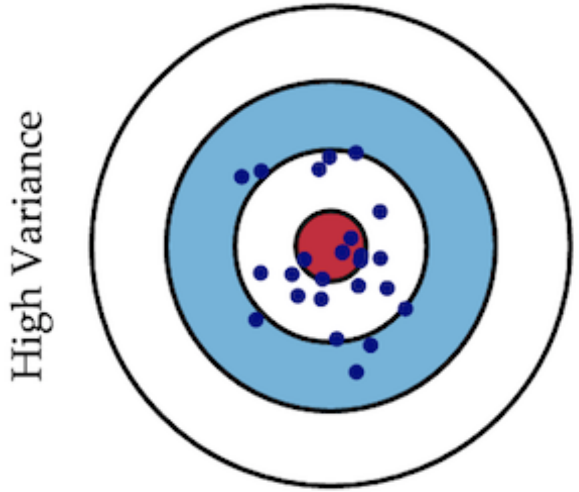


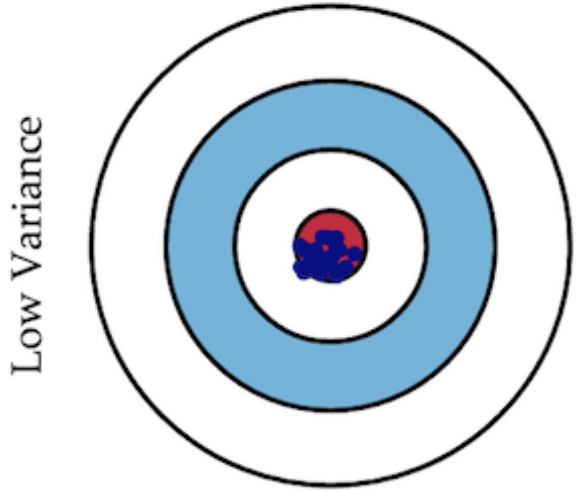
BIAS VS. VARIANCE

- The Variance is how much the predictions for a given point vary between different realisations of the model

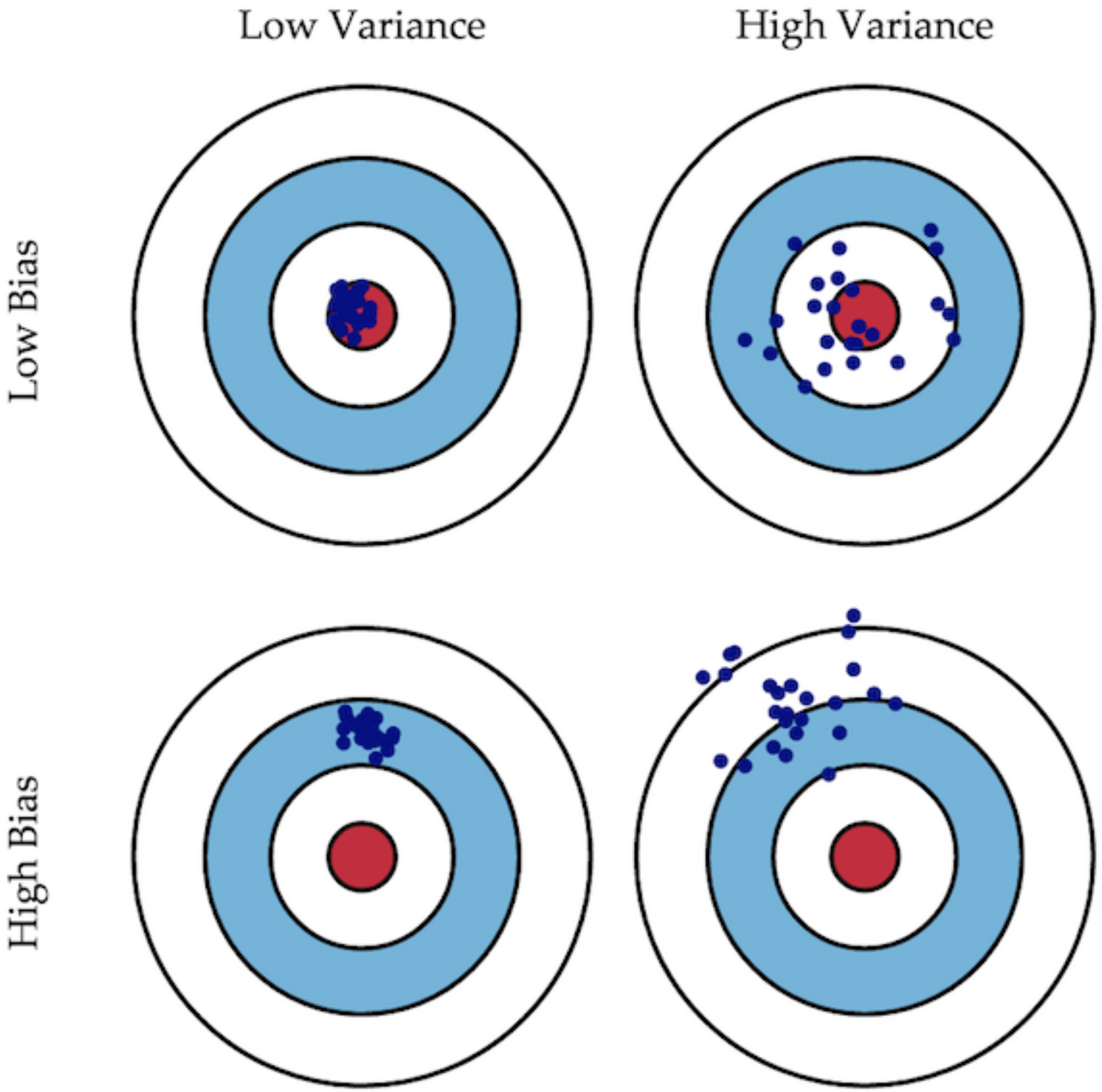
- Error due to Variance is taken as the variability of a model prediction for a given point

- Imagine creating multiple models on various datasets

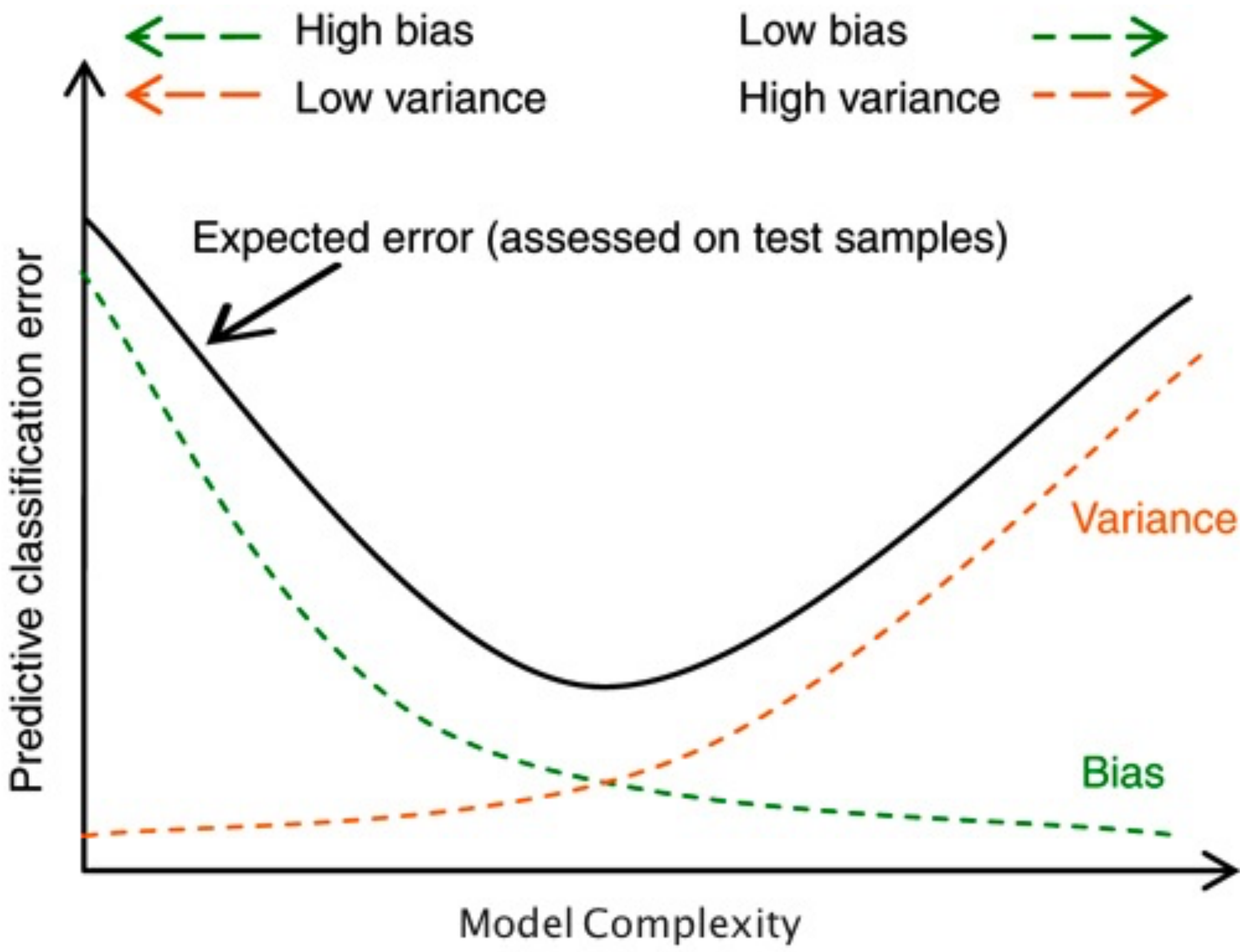




BIAS VS. VARIANCE



BIAS VS. VARIANCE



VARIANCE

- Variance measures how far a dataset is spread out. It can vary
  - from 0, meaning no Variance (all the data is the same)
  - to theoretically infinity, meaning there are differences between the data
- Its unit is the square of the variable!
- Variance of the Population

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

- **Unbiased** Variance of the Sample
  - Note the (n-1)

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)}$$

STANDARD DEVIATION

- Standard Deviation 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 the Variance
- Standard Deviation of the Population $$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$
- Standard Deviation of the Sample $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}}$$

MEAN STANDARD ERROR

- The **Mean Standard Error** quantifies the precision of the Mean

- It is a measure of how far the Sample Mean is likely to be from the True Population Mean

- It generally increases with the size of an estimate, meaning a large Standard Error may not indicate the estimate of the mean is unreliable

- It is often better to compare the error in relation to the size of the estimate

MEAN STANDARD ERROR

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- There are other Standard Errors, such
  - Standard error of proportion
  - Standard error of difference for proportions
  - Standard error of difference of sample proportions
  - Standard error of difference of sample means
  - Standard error of difference of paired sample means

CODE ALONG PART 3: VARIANCE AND STANDARD DEVIATION

- You can calculate Variance and Standard Deviation easily in Pandas

- Methods available include

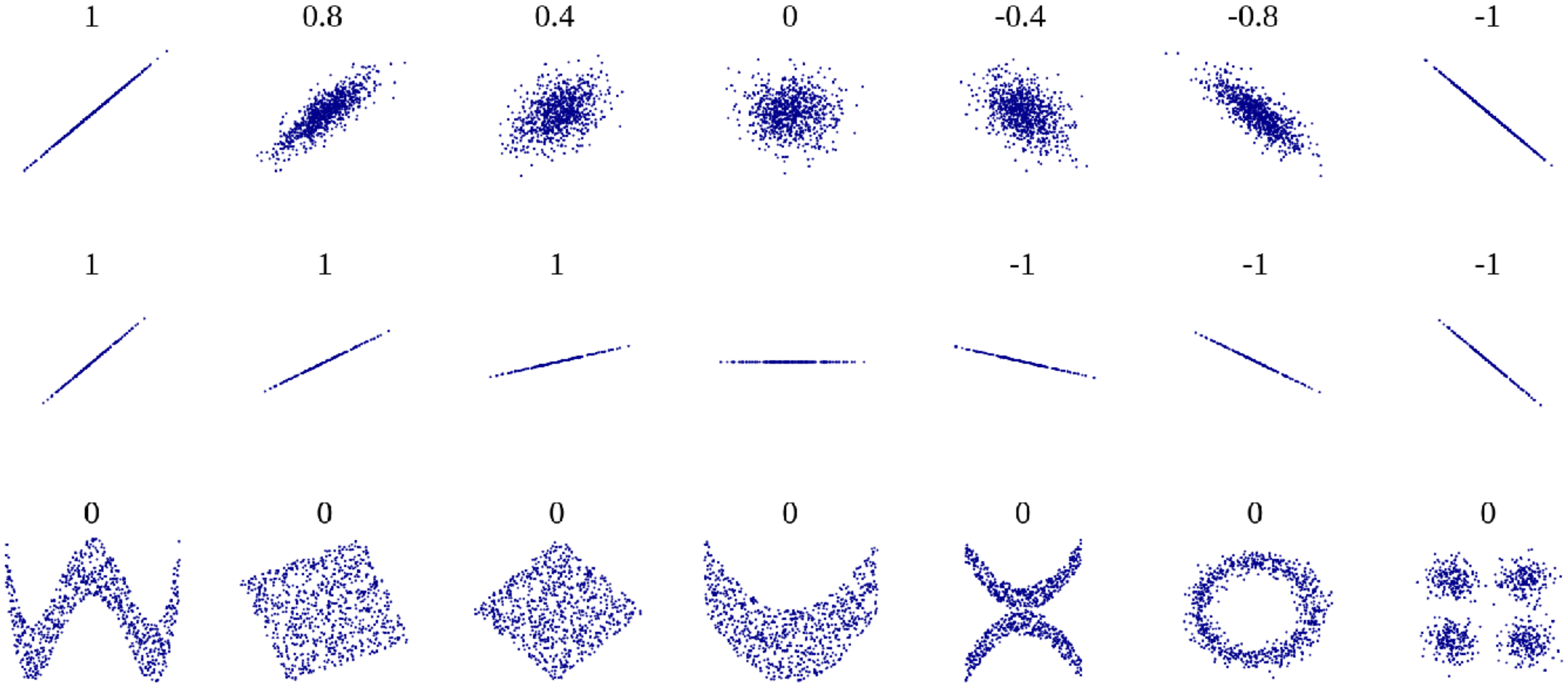
  .var() - Compute Variance

  .std() - Compute Standard Deviation

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

CORRELATION

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



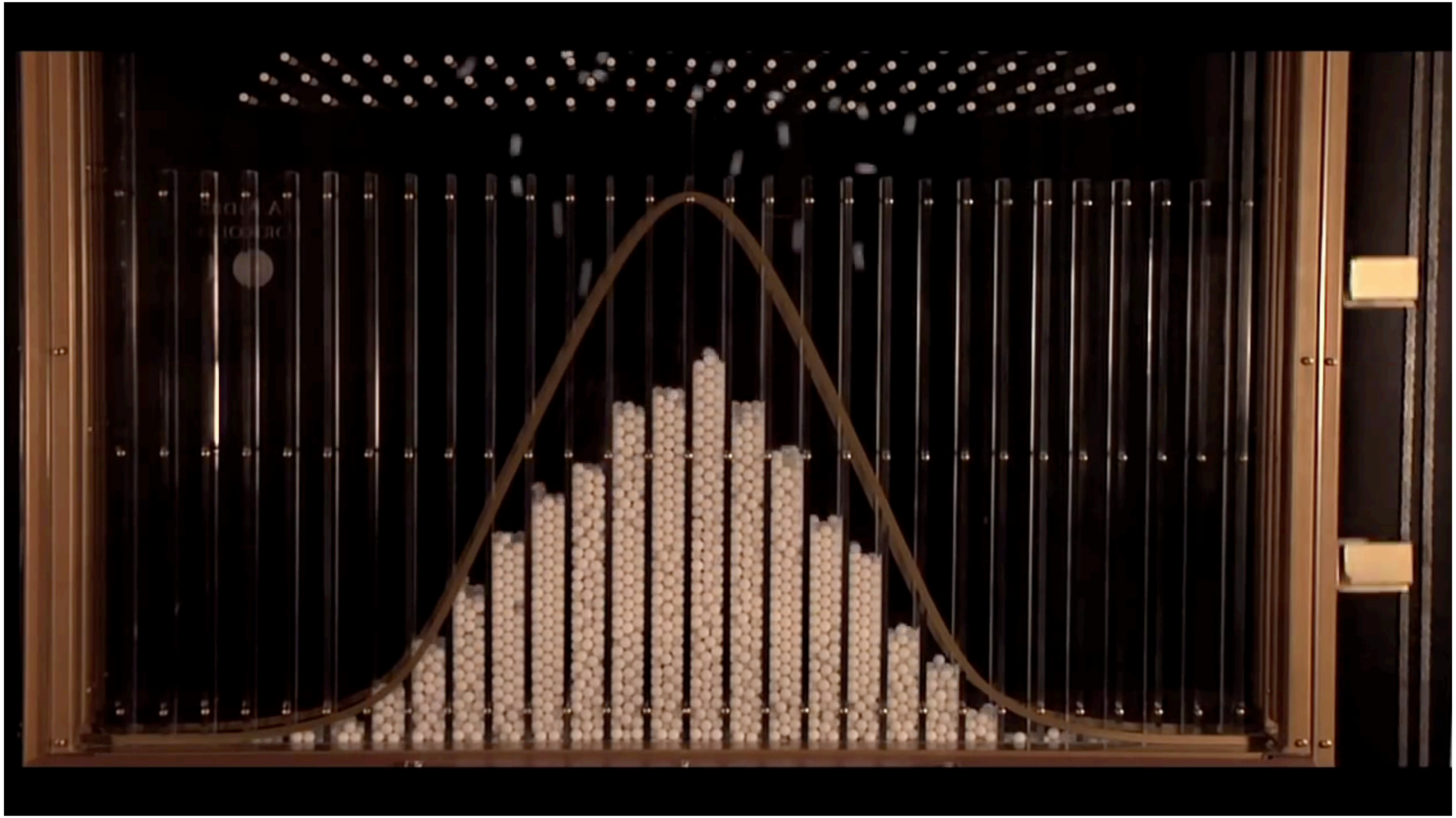
CONTEXT

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

INTRODUCTION

IS THIS NORMAL?

IS THIS NORMAL?



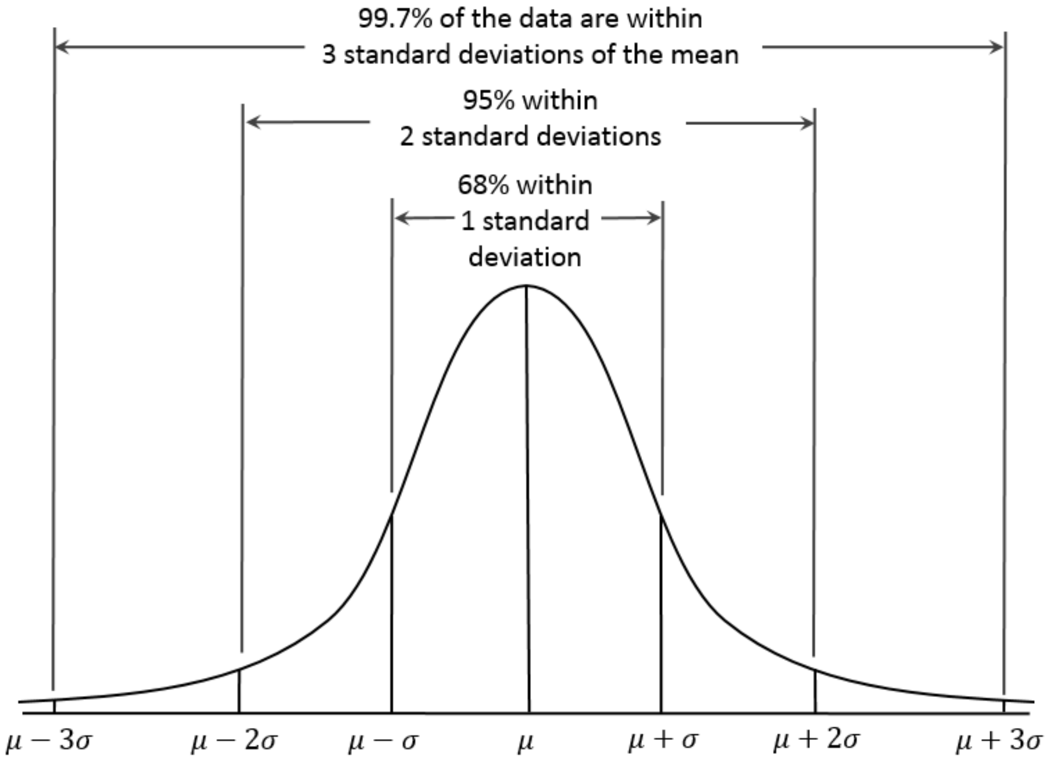
THE NORMAL DISTRIBUTION

- A Normal distribution is often a key assumption to many models

- The Normal distribution is a function of the Mean and the Standard Deviation

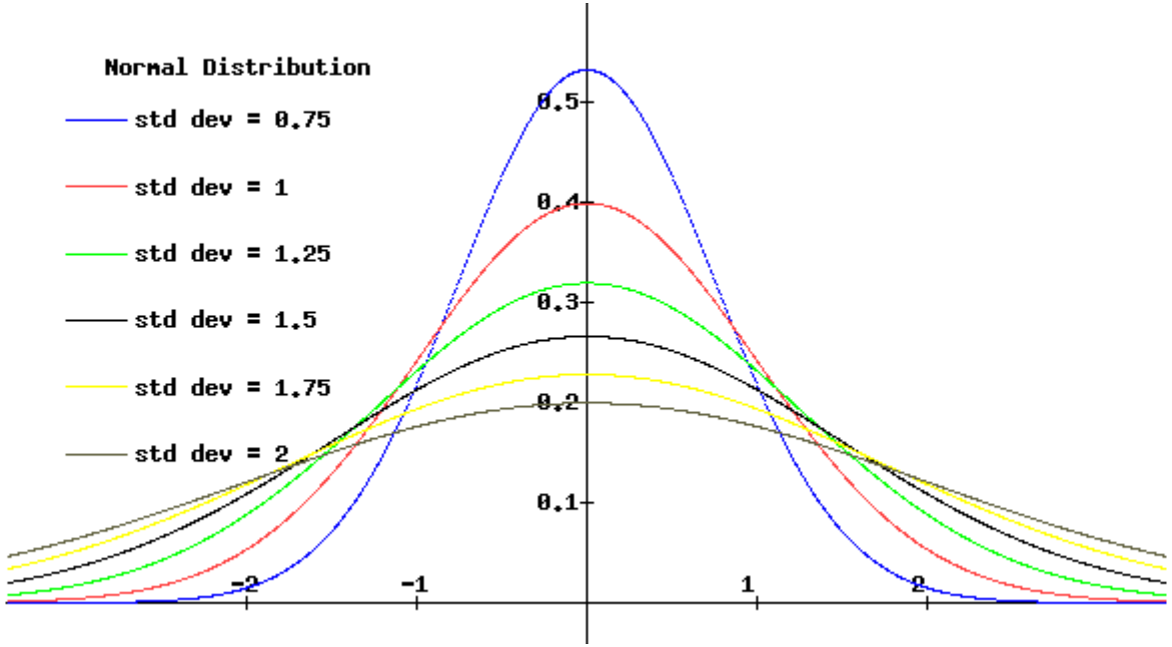
  - The Mean determines the centre of the distribution

  - The Standard Deviation determines the height and width of the distribution



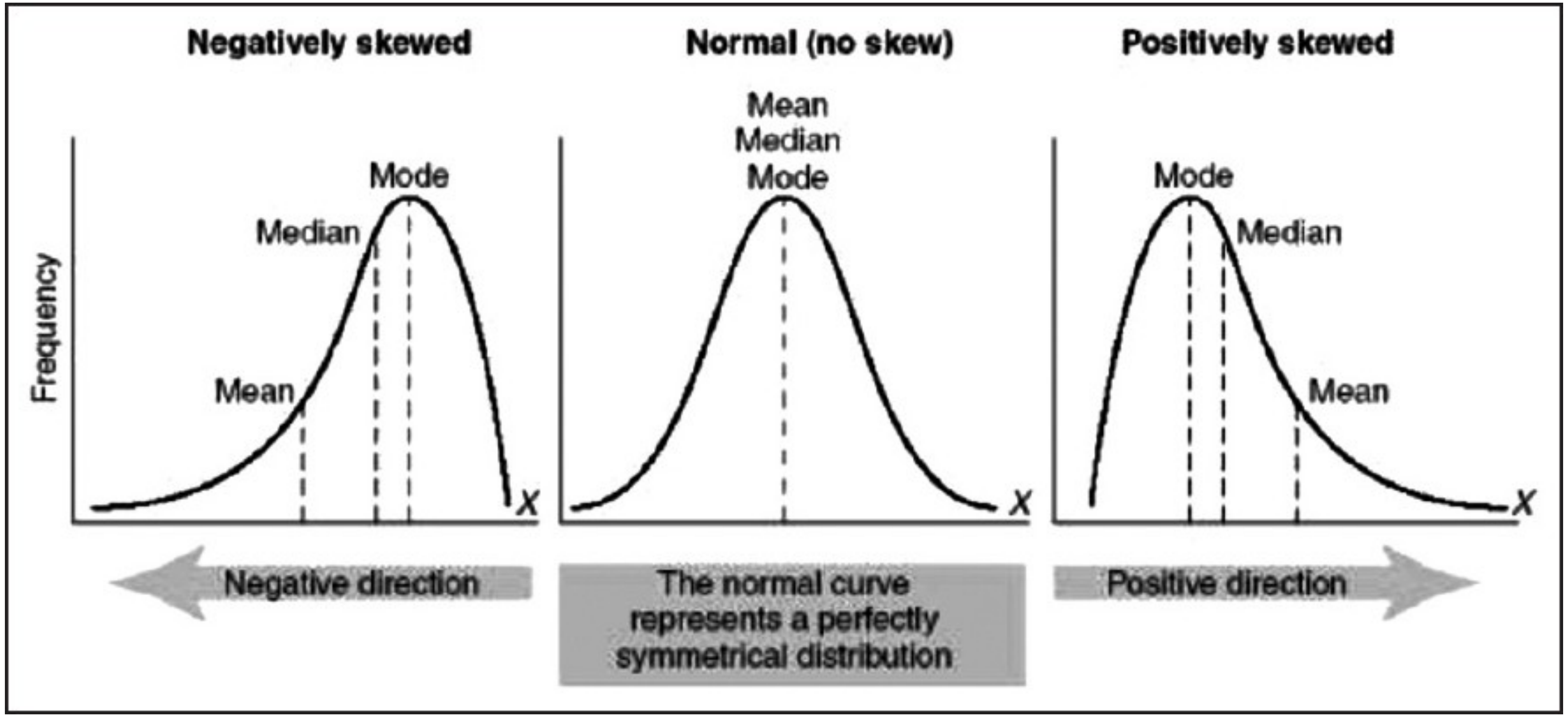
THE NORMAL DISTRIBUTION

- 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 it tall and narrow



SKEWNESS

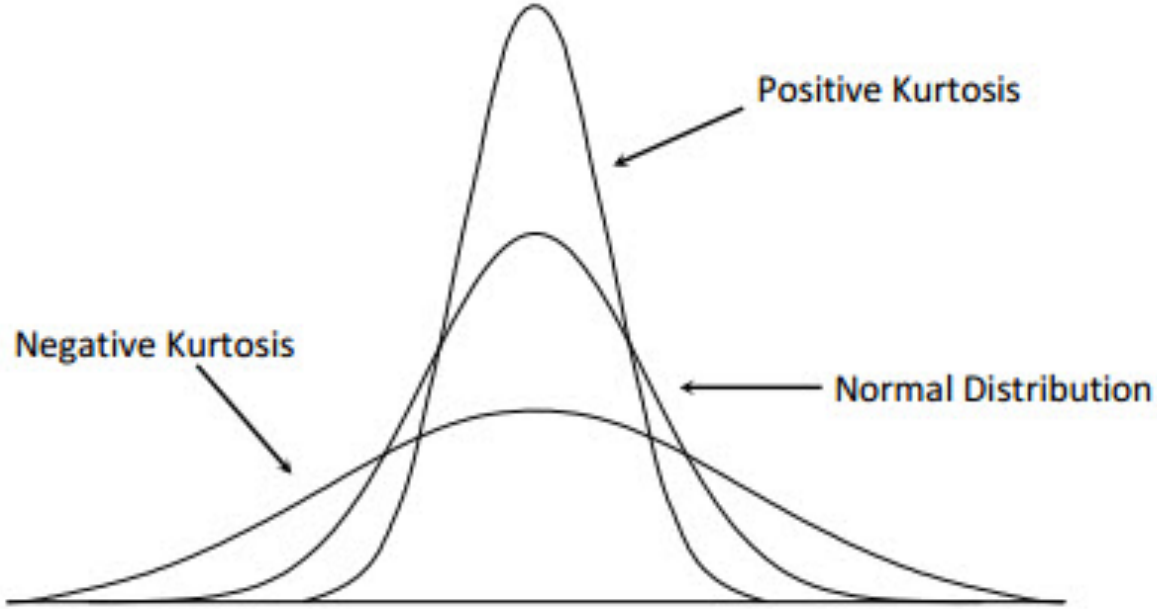
- Skewness is a measure of the asymmetry of the distribution of a random variable about its mean
- Skewness can be positive, negative or even undefined



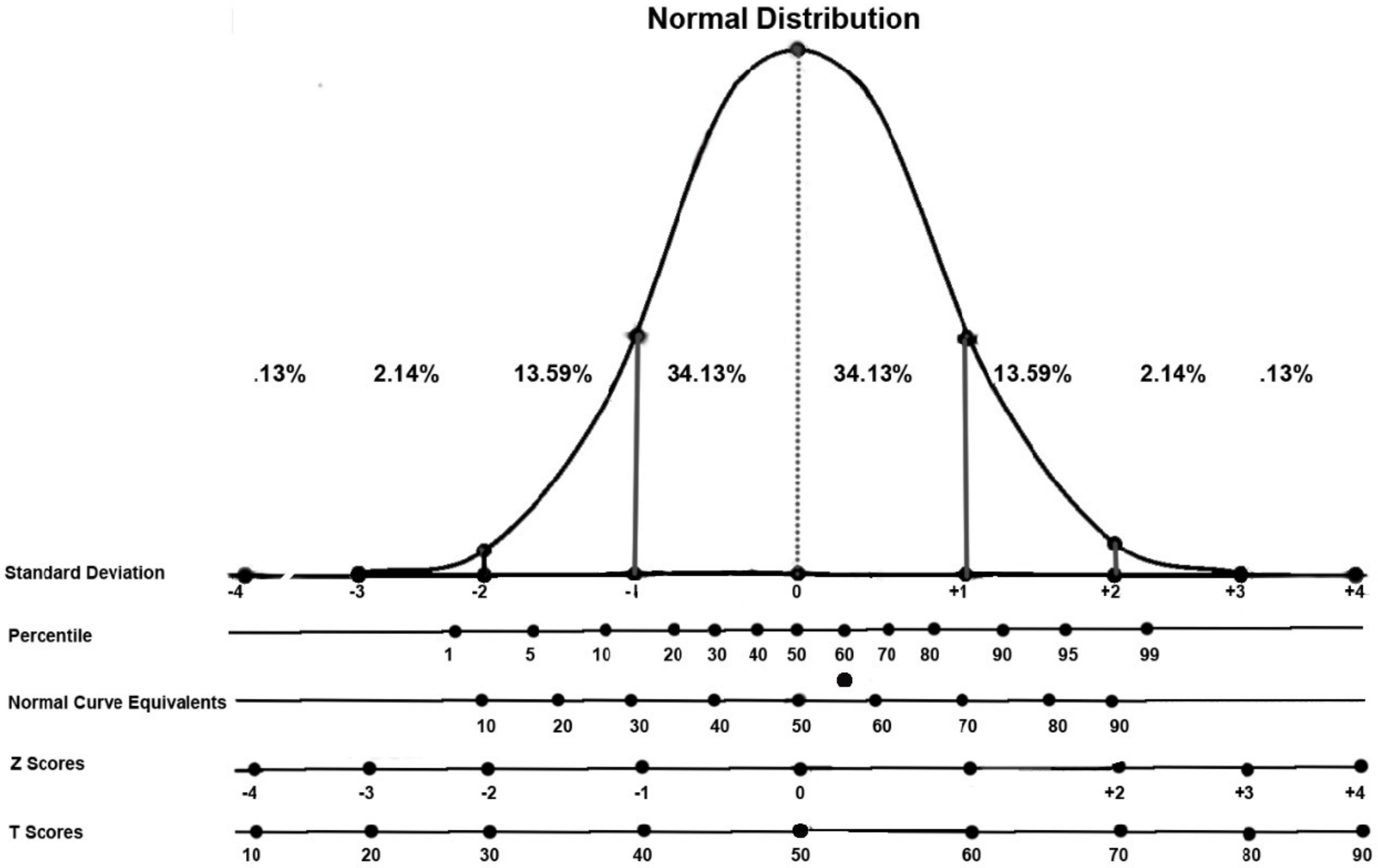
KURTOSIS

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

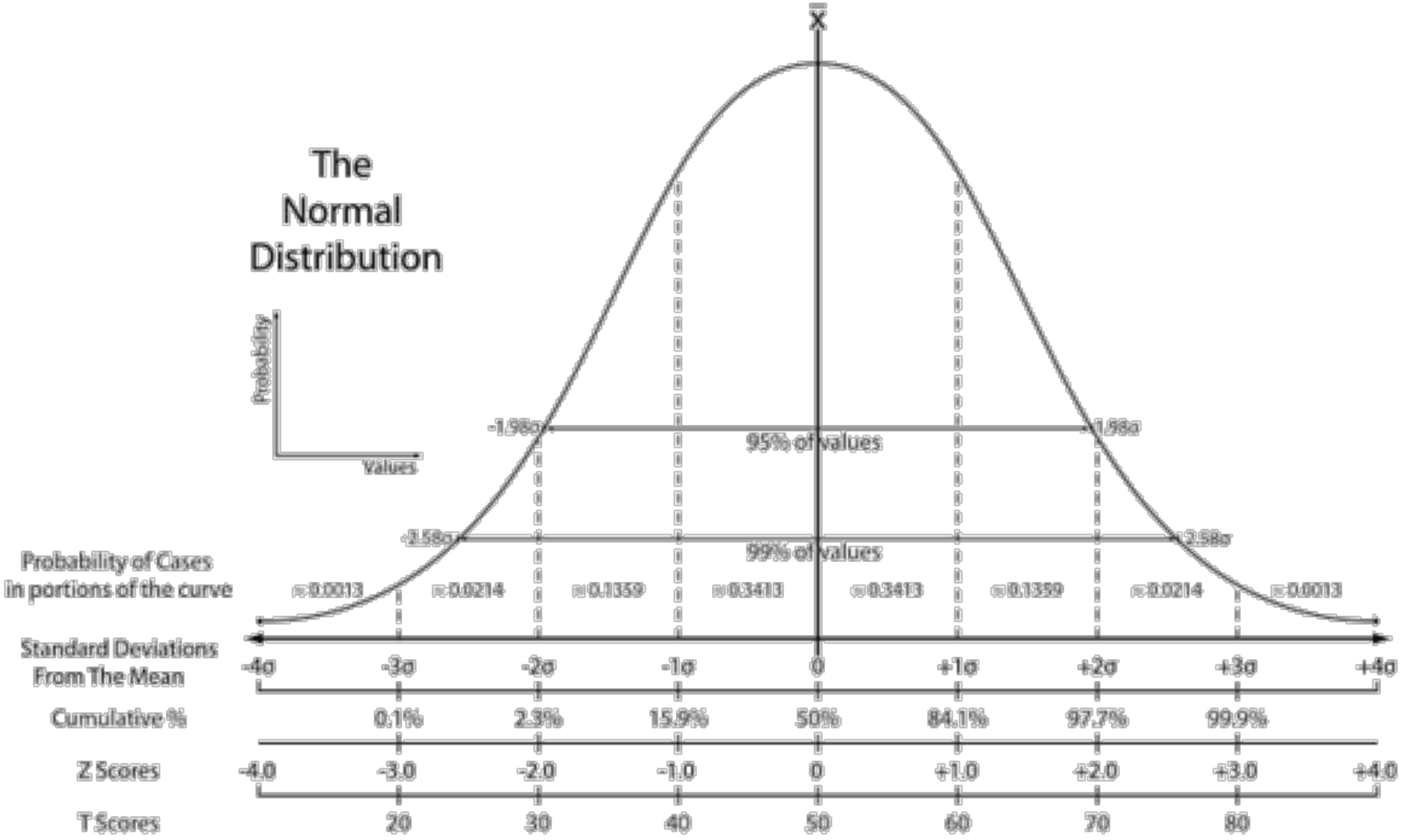
- Datasets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly and have heavy tails



THE NORMAL DISTRIBUTION



THE NORMAL DISTRIBUTION



DEMONSTRATION

DETERMINING THE DISTRIBUTION OF THE DATA

CODE ALONG: DETERMINING THE DISTRIBUTION OF THE DATA

- Open the starter code notebook located at
  - ~/lessons/lesson-03/code/lesson-3-demo.ipynb
- Ask your classmates and instructor for help if you have problems!

GUIDED PRACTICE

IS THIS SKEWED?

ACTIVITY: IS THIS SKEWED?

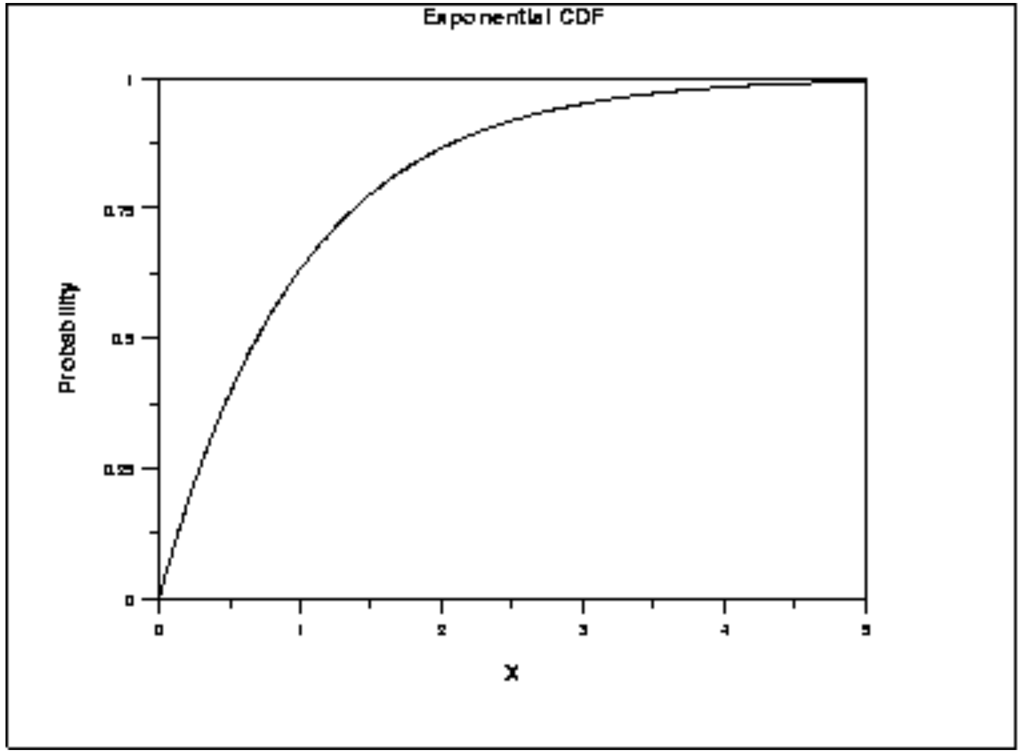
EXERCISE

DIRECTIONS (10 MINUTES)

1. We are going to walk through several images of datasets
2. For each image, vote on whether the distribution is
   a. Normal
   b. Positively, negatively or not skewed
   c. Has positive, negative or zero kurtosis
3. Determine how you would correct the issue with each dataset to return it to the normal distribution

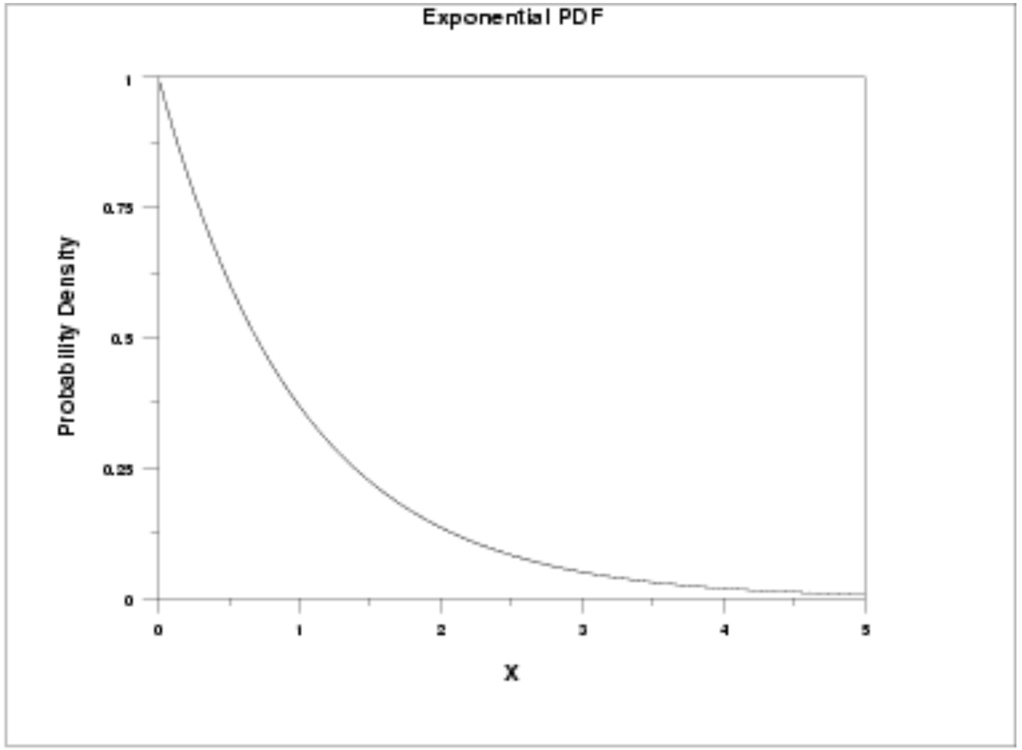
ACTIVITY: IS THIS SKEWED?

- Is the distribution Normal?
- Positively, negatively or not skewed
- Positive, negative or zero kurtosis
- How to converge to a Normal distribution (if possible)



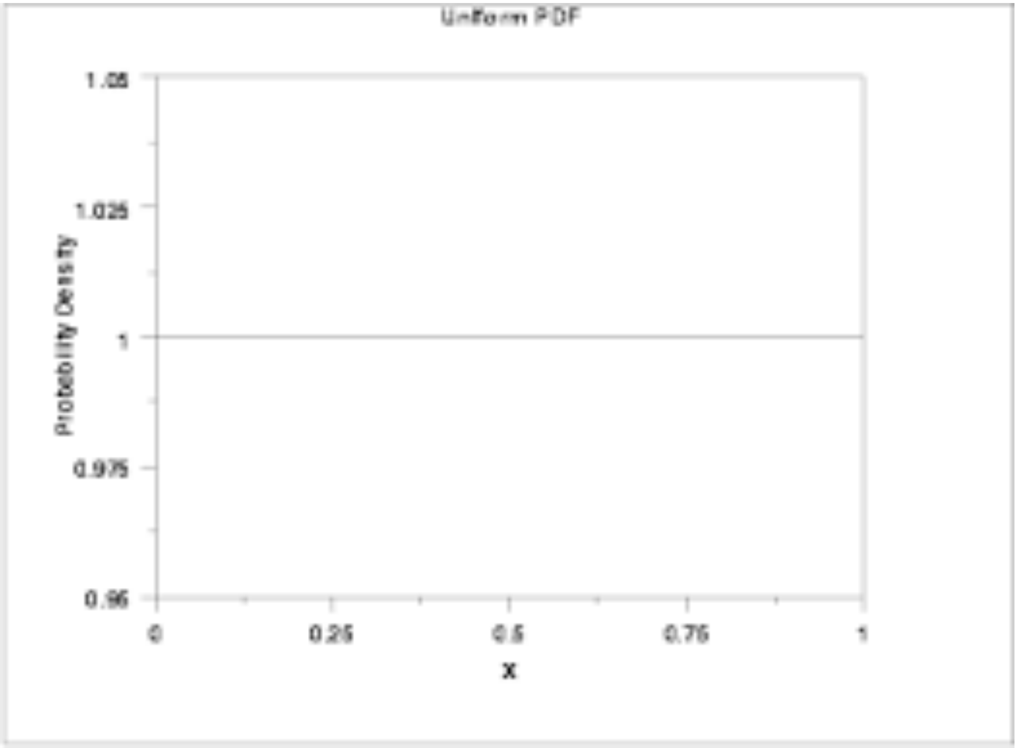
ACTIVITY: IS THIS SKEWED?

- Is the distribution Normal?
- Positively, negatively or not skewed
- Positive, negative or zero kurtosis
- How to converge to a Normal distribution (if possible)



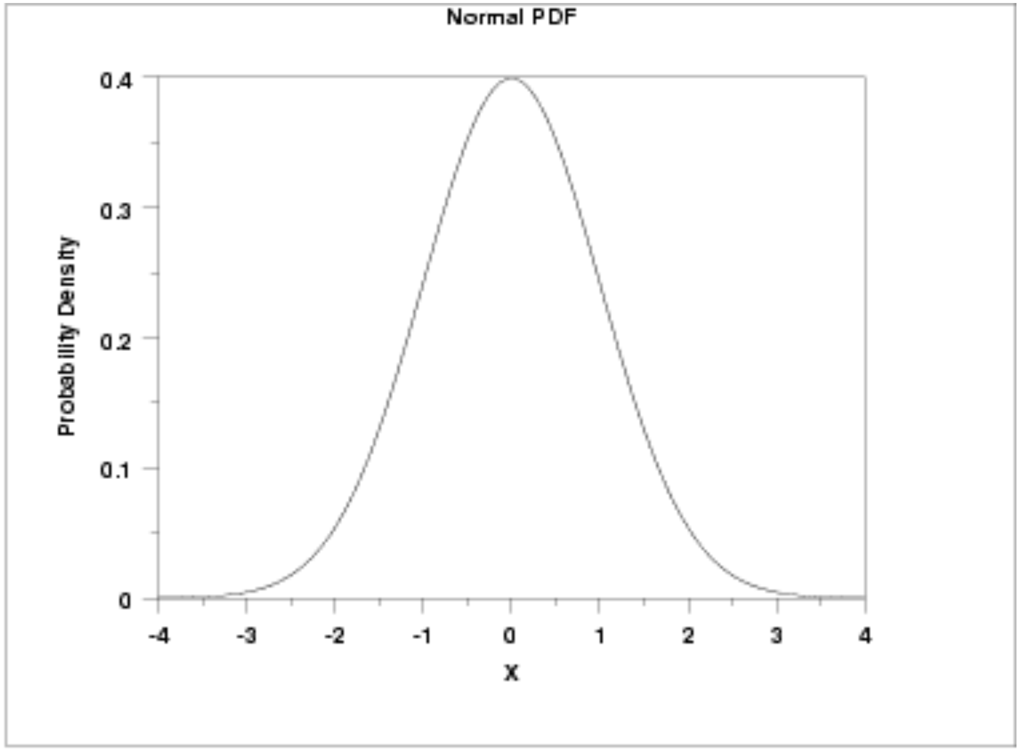
ACTIVITY: IS THIS SKEWED?

- Is the distribution Normal?
- Positively, negatively or not skewed
- Positive, negative or zero kurtosis
- How to converge to a Normal distribution (if possible)



ACTIVITY: IS THIS SKEWED?

- Is the distribution Normal?
- Positively, negatively or not skewed
- Positive, negative or zero kurtosis
- How to converge to a Normal distribution (if possible)



INTRODUCTION

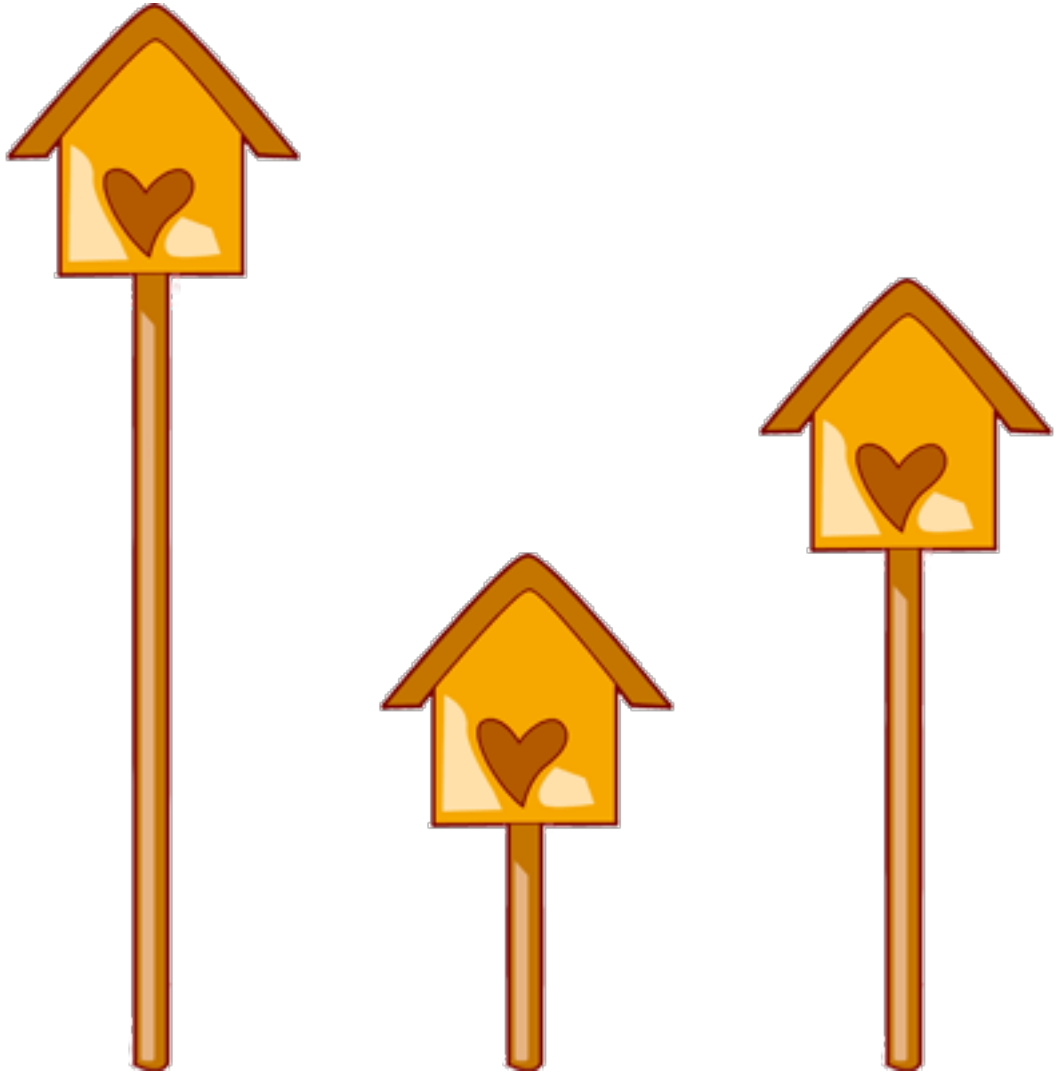
VARIABLE TYPES

VARIABLE TYPES

- Numeric variables can take on a large range of non-predetermined, quantitative values
  - These are things such as height, income

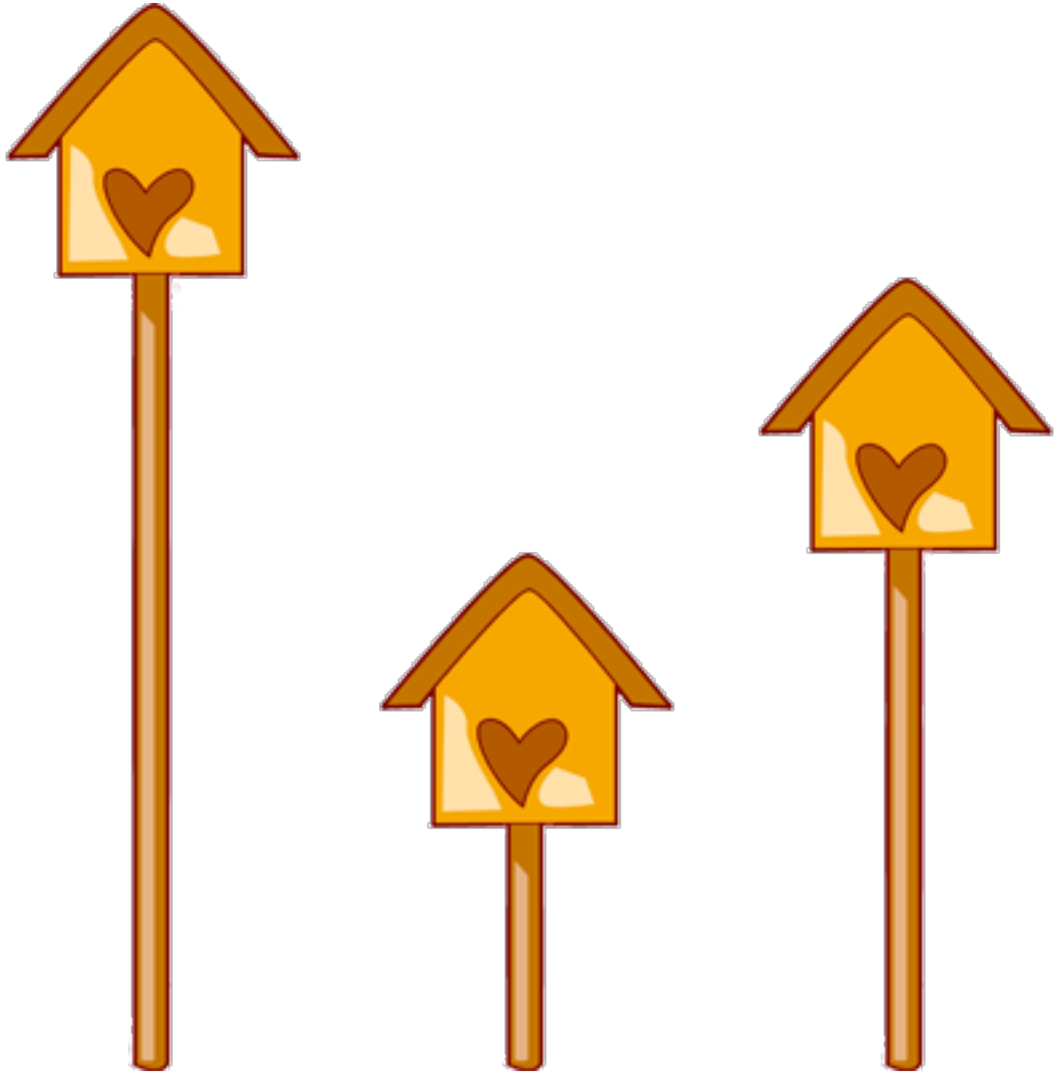
- Categorical variables can take on a specific set of variables
  - These are things such as race, gender, paint colours, movie titles

MEASUREMENTS



MEASUREMENTS

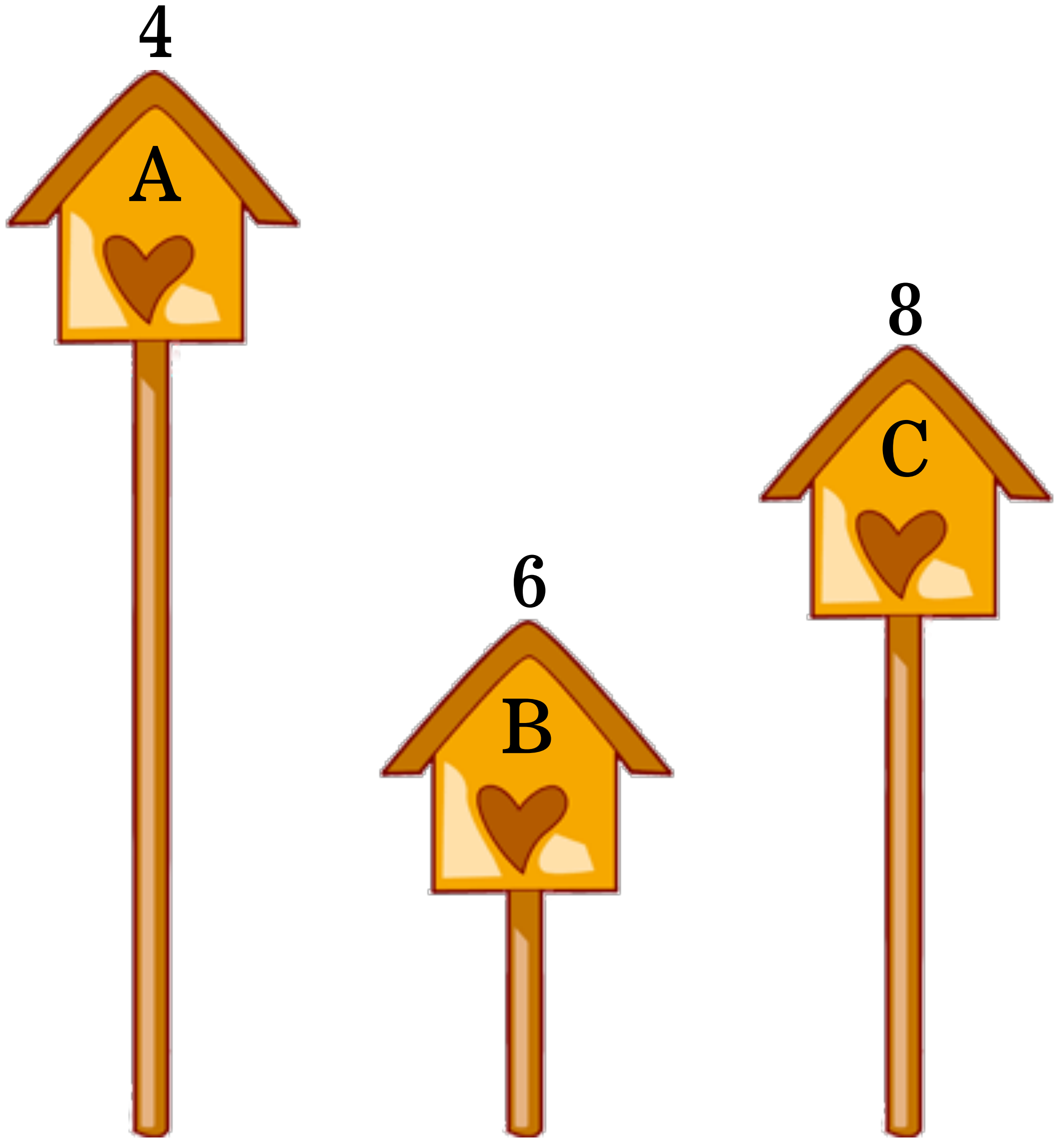
- Measurement is the representation of relations between objects, persons or groups on a certain property by using relations between numbers



MEASUREMENTS

- Measurement is the representation of relations between objects, persons or groups on a certain property by using relations between numbers

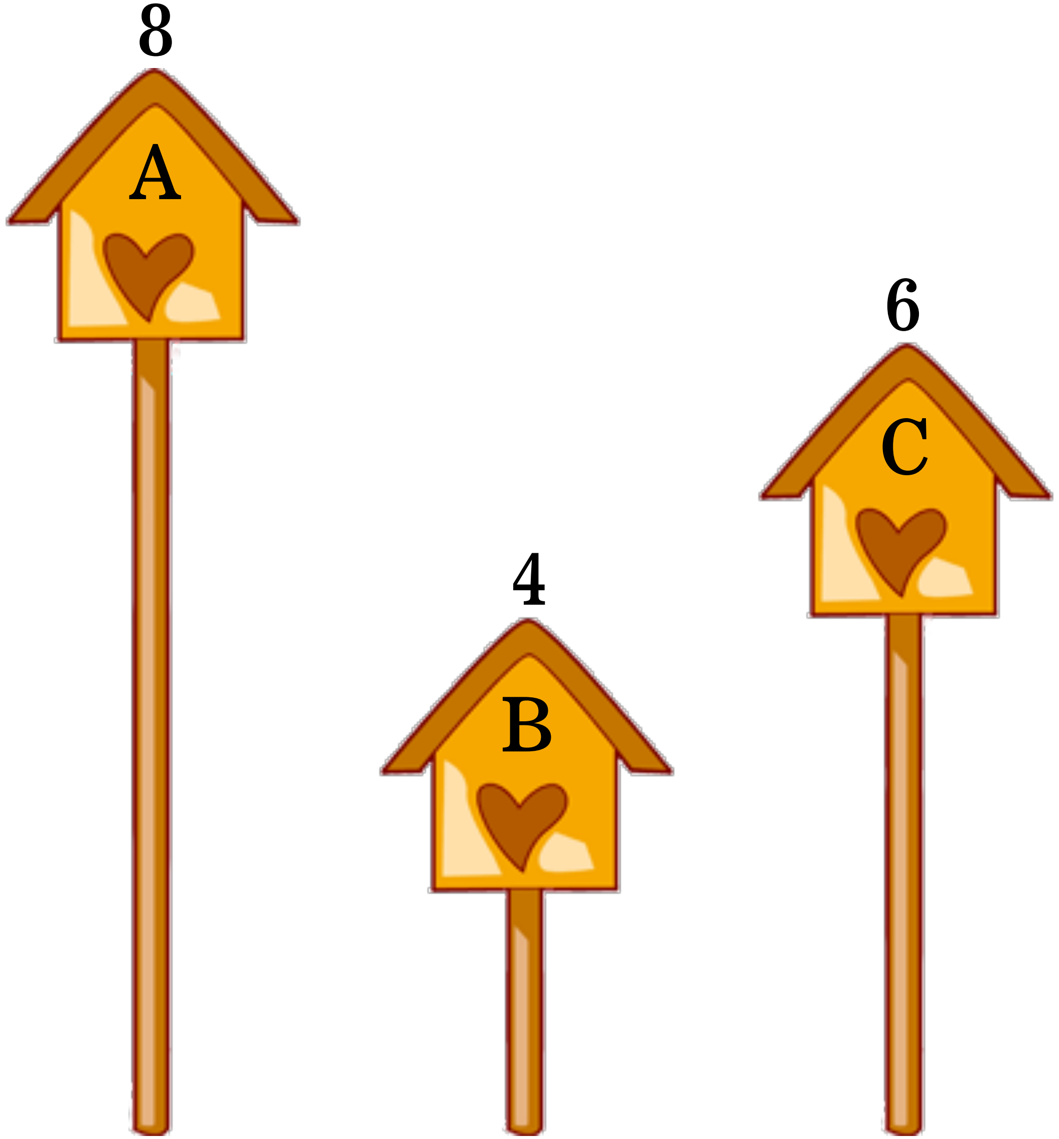
- Inequality: $(A \neq B \neq C)$ or $(4 \neq 6 \neq 8)$



MEASUREMENTS

- Measurement is the representation of relations between objects, persons or groups on a certain property by using relations between numbers

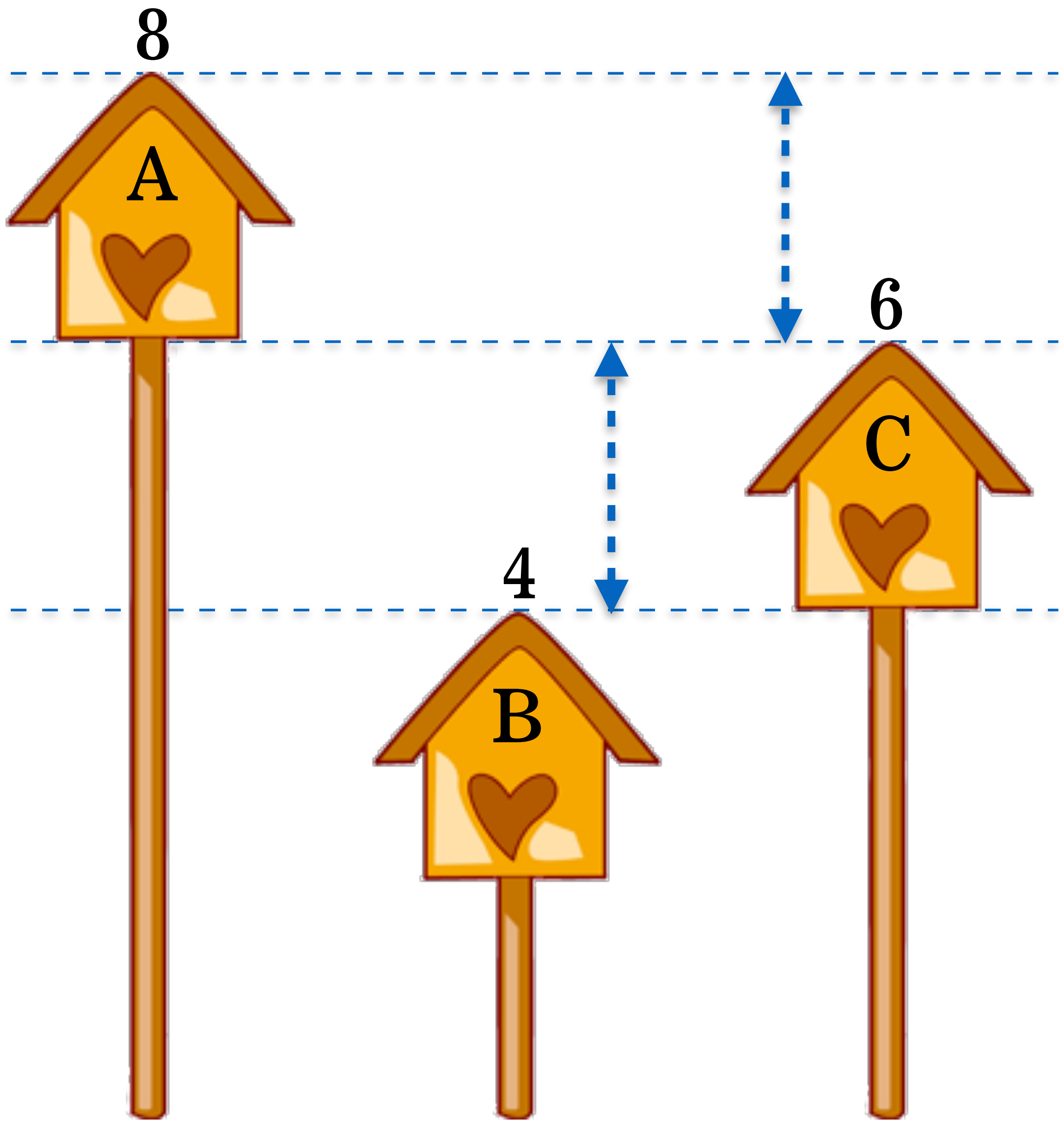
- Inequality: (A ≠ B ≠ C) or (4 ≠ 6 ≠ 8)
- Order: (A > B < C) or (8 > 4 < 6)



MEASUREMENTS

- Measurement is the representation of relations between objects, persons or groups on a certain property by using relations between numbers

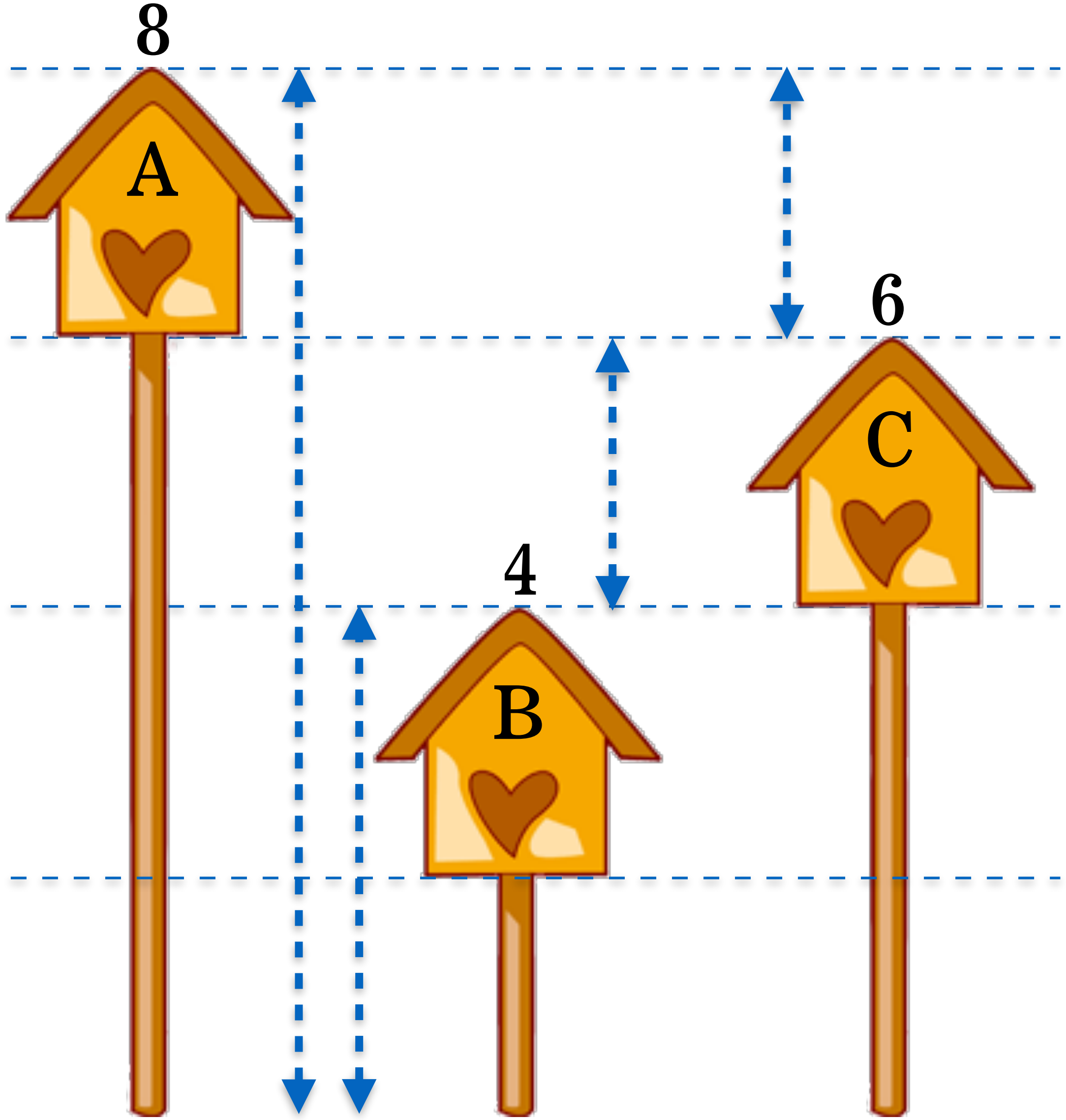
- Inequality: $(A \neq B \neq C)$ or $(4 \neq 6 \neq 8)$
- Order: $(A > B < C)$ or $(8 > 4 < 6)$
- Difference: $(A - C) = (C - B) = 2$



MEASUREMENTS

- Measurement is the representation of relations between objects, persons or groups on a certain property by using relations between numbers

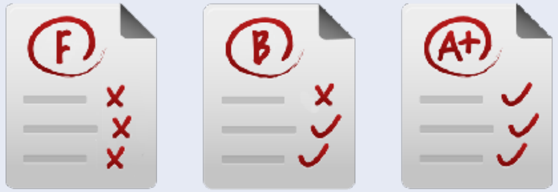
- Inequality: $(A \neq B \neq C)$ or $(4 \neq 6 \neq 8)$
- Order: $(A > B < C)$ or $(8 > 4 < 6)$
- Difference: $(A - C) = (C - B) = 2$
- Ratio: $(A = 2 \cdot B)$ or $(8 = 2 \cdot 4)$



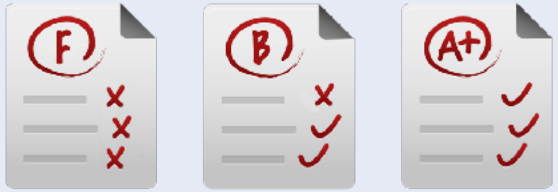
MEASUREMENT LEVELS

Relation	Level	Comment	Example
Inequality	Nominal	qualitative, descriptive, categories	

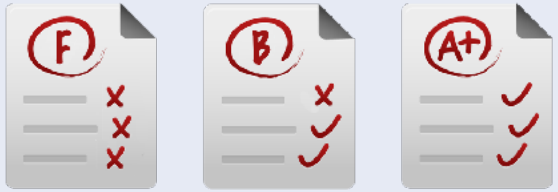
MEASUREMENT LEVELS

Relation	Level	Comment	Example
Inequality	Nominal	qualitative, descriptive, categories	
Order	Ordinal	ordering or ranking (no information about distance between ranks)	

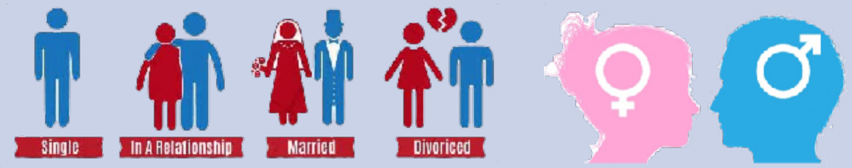
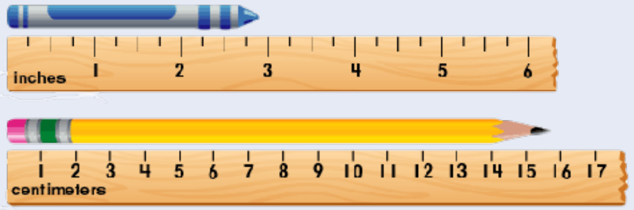
MEASUREMENT LEVELS

Relation	Level	Comment	Example
Inequality	Nominal	qualitative, descriptive, categories	
Order	Ordinal	ordering or ranking (no information about distance between ranks)	
Difference	Interval	arbitrary or **no** natural zero (zero is a meaningful value)	-40 -30 -20 -10 0 10 20 30 40 50 °C -40 -20 0 20 40 60 80 100 120 °F

MEASUREMENT LEVELS

Relation	Level	Comment	Example
Inequality	Nominal	qualitative, descriptive, categories	
Order	Ordinal	ordering or ranking (no information about distance between ranks)	
Difference	Interval	arbitrary or **no** natural zero (zero is a meaningful value)	°C -40 -30 -20 -10 0 10 20 30 40 50 °F -40 -20 0 20 40 60 80 100 120
Ratio	Ratio	**no** arbitrary or natural zero (zero means an absence of a value)	

VARIABLE TYPES

Categorical	**Unordered**	Binary (only two values) or more than two	Marital Status Gender	
	Ordered		Satisfaction	
Numerical	**Discrete**	Count	Number of children	
	Continuous	Measure	Length	

DEMONSTRATION

CLASSES

CLASS / DUMMY VARIABLES

- Let’s say we have the categorical variable area, which takes on one of the following values
  - rural
  - suburban
  - 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
  - 2 = urban

CLASS / DUMMY VARIABLES

- 0 = rural, 1 = suburban, 2 = urban

- But this implies an ordered relationship
  - Is urban twice suburban?
  - That does not make sense

- However, we can represent this information by converting the one area variable into two new variables
  - area_urban
  - area_suburban

CLASS / DUMMY VARIABLES

- We will 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 is 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 will choose rural as our reference category

- Step 2: Convert the values rural, 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

- 0 = rural, 1 = suburban, 2 = urban

- Why do we need only two dummy variables?

- We can derive all of the possible values from these two
  - If an area is not 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 area_urban=0 and area_suburban=0
    - then the area must be rural

CLASS / DUMMY VARIABLES

- We can do this for a gender variable with two categories
  - male
  - female

- How many dummy variables need to be created?

CLASS / DUMMY VARIABLES

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

ACTIVITY: DUMMY COLOURS

EXERCISE

DIRECTIONS (15 MINUTES)

1. It is important to understand the concept before we use the Pandas function get_dummies to create dummy variables. So today, we will create our dummy variables by hand

   a. Draw a table like the one on the white board

   b. Create dummy variables for the variable colours that has six categories

      i. blue, red, green, purple, grey and brown

      ii. Use grey as the reference

REVIEW

CONCLUSION

- Let’s go through the process for creating dummy variables for colours

  - We talked about several different types of summary statistics
    - what are they?
  - We covered different types of visualisations
    - which ones?
  - We talked about the normal distribution
    - how do we determine the distribution of the data?

- Any other questions?

BEFORE NEXT CLASS

DUE DATE

- Project: Unit Project 2
  - ~/projects/unit-projects/project-02/README.md

Q & A

CREDITS AND REFERENCES

STATISTICS FUNDAMENTALS I

- The Mathematics of Love - TED Talk
  - Hannah Fry
    “Mathematician, Science Presenter and All Round Badass”
    - Hannah Fry researches the trends in our civilisation and ways we can forecast its future
      - Personal Website
      - The Mathematics of Love - TED Talk
- From Chaos to Order on the Galton Machine
  - youTube video