Descriptive Statistics for Exploratory Data Analysis

Ivan Corneillet

Data Scientist

GENERAL ASSEMBLY

Learning Objectives

After this lesson, you should be able to:

- Identify variable types
- Use the *pandas* and *NumPy* libraries to analyze datasets using basic summary statistics: mean, median, mode, max, min, quartile, inter-quartile range, variance, standard deviation, and correlation
- Create data visualizations – including boxplots, histograms, and scatter plots – to discern characteristics and trends in a dataset

Announcements and Exit Tickets

Review

Review

❶ *IDENTIFY the problem*

The SMART Framework for Data Science

❶ IDENTIFY the problem | The SMART Framework for Data Science:

SPECIFIC	The dataset and key variables are clearly defined
MEASURABLE	The type of analysis and major assumptions are articulated
ATTAINABLE	The question you are asking is feasible for your dataset and is not likely to be biased
REPRODUCIBLE	Another person (or you in 6 months!) can read your state and understand exactly how your analysis is performed
TIME-BOUND	You clearly state the time period and population for which this analysis will pertain

Trends often change over time and vary by the population of source of your data. It is important to clearly define who/what you included in your analysis as well as the time period for the analysis

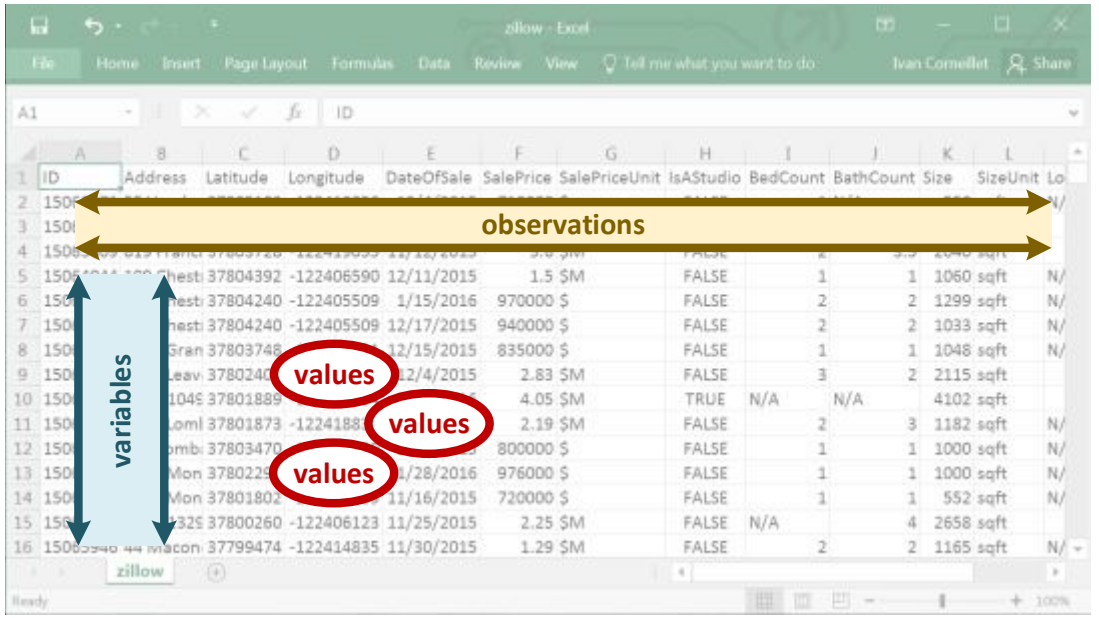
Review

❸ *Parse the Data*

Tidy Data and pandas

❸ PARSE the Data | Tidy Data: a tabular format suitable for *pandas* and machine learning algorithms

- The three rules of tidy data:
  - Each observation is placed in its own row
  - Each variable in the dataset is placed in its own column
  - Each value is placed in its own cell



Activity | Subsetting with *pandas* (10 minutes)

EXERCISE

	DataFrame	Series
Column subsetting		
by name		
by location		
Row subsetting		
by index label		
by location		
Cell subsetting/scalar lookup		
By index label/column name		
By location		

	DataFrame	Series
Column subsetting		

by name
 	
(Columns names are stored in `df.columns`)
(`df.columns.get_loc('X1')` returns X1's column index)
	`# New DataFrame with column named X1`
	`df[ ['X1'] ]`
	
	`# 2+ columns (in the order listed)`
	`df[ ['X1', 'X2', …] ]`
		`df['X1']`
		
		`df.X1`

by location	`# New DataFrame with column at location i (numbering starts at 0)`
	`df[ [column_i] ]`
	
	`# 2+ columns (in the order listed)`
	`df[ [column_i, column_j, …] ]`	

Row subsetting		

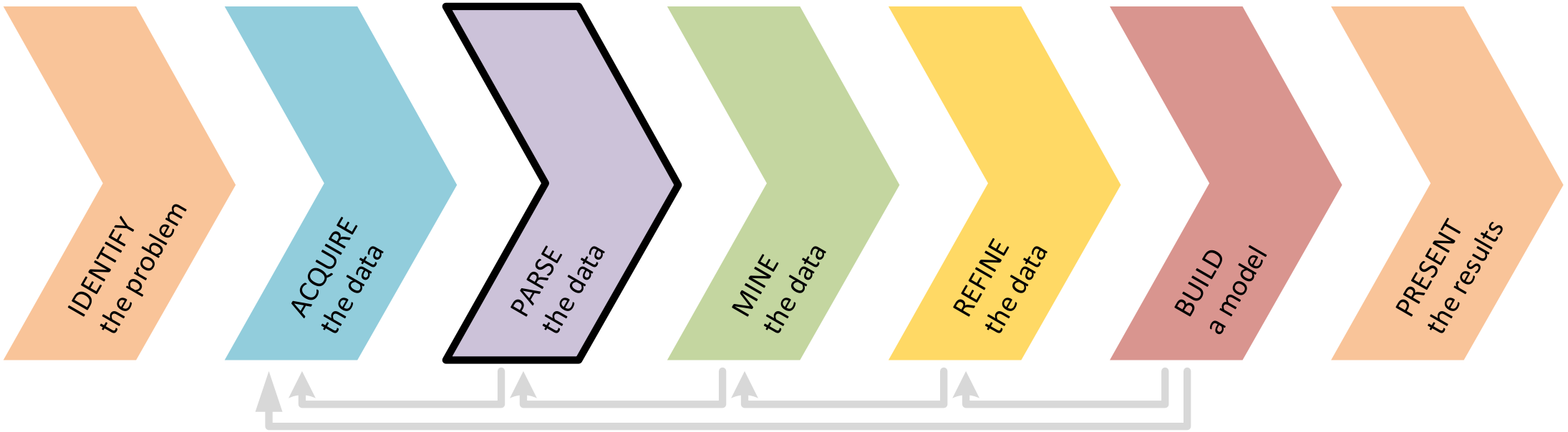
by index label	`df.loc[ [index_label_i] ]`
	`df.loc[ [index_label_i, index_label_j, …] ]`
	
	`# Can use a range if the index is made of numbers (rows "a" to "b" included)`
	`df.loc[ index_label_a : index_label_b ]`
		`df.loc[index_label_i]`

by location	`df.iloc[ [row_i] ]`
	`df.iloc[ [row_i, row_j, …] ]`
	
	`# (rows "a" to "b' excluded)`
	`df.iloc[row_a : row_b ] or df[row_a : row_b ]`
		`df.iloc[location_i]`

Cell subsetting/scalar lookup		
By index label/column name	`df.at[index_label, 'X1']`	
By location	`df.iat[row_i, column_j]`	

Today

Today we'll keep our focus on PARSE the data



Today, we are covering Research Design and introducing the *pandas* library

Research Design and Data Analysis	Research Design	Data Visualization in *pandas*	Statistics	Exploratory Data Analysis in *pandas*
Foundations of Modeling	Linear Regression	Classification Models	Evaluating Model Fit	Presenting Insights from Data Models
Data Science in the Real World	Decision Trees and Random Forests	Time Series Data	Natural Language Processing	Databases

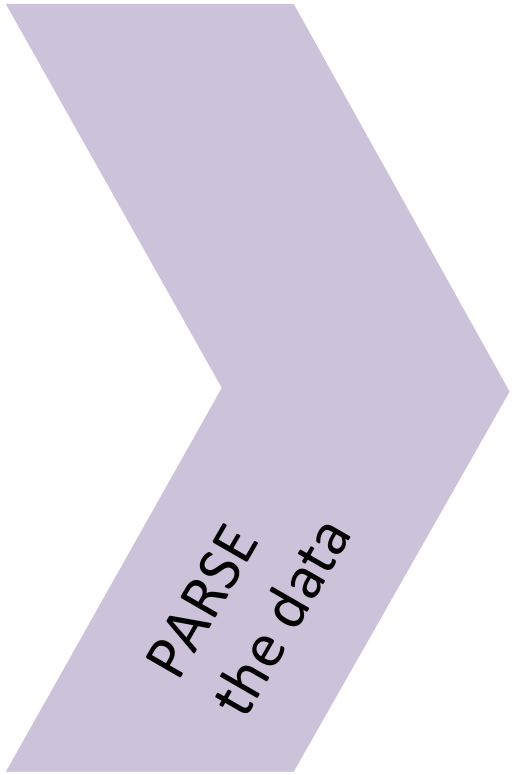
Here's what's happening today:

- Announcements and Exit Tickets
- Review
- ❸ Parse the Data
  - Types of Data and Types of Measurement Scales
  - Populations and Samples; Descriptive vs. Inferential Statistics
  - Measures of Central Tendency and Measures of Dispersion
  - Boxplots
  - Outliers
  - Histograms
  - Correlation
- Review
- Exit Tickets

Q & A

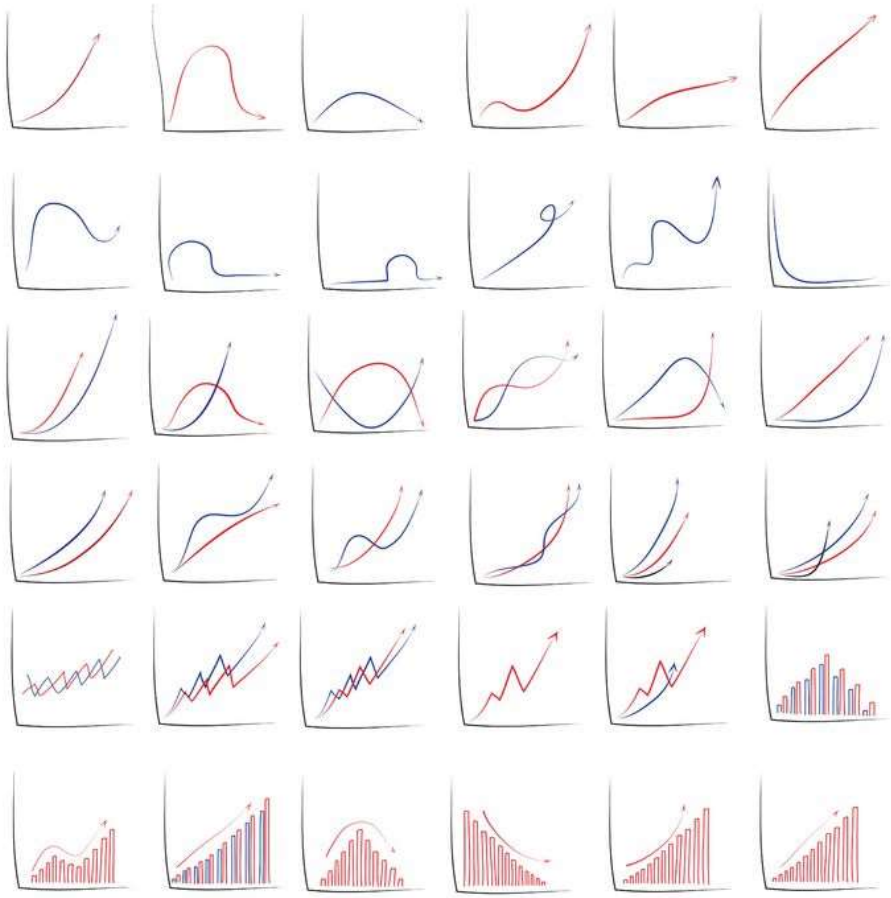
❸ PARSE the Data

❸ Parse the Data



- Parse the Data
  - *Read any documentation provided with the data (session 2)*
  - **Perform exploratory data analysis (session 3)**
    - *Verify the quality of the data (sessions 2/3)*

The main theme today is to have enough statistics knowledge to perform Exploratory Data Analysis



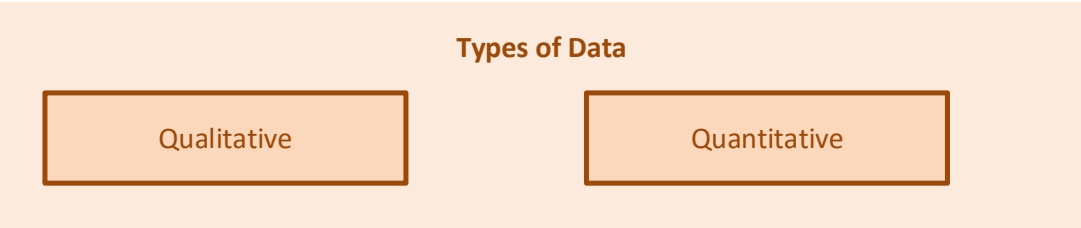
Napat Polchoke © 123RF.com

- Types of Data and Types of Measurement Scales
- Populations and Samples; Descriptive vs. Inferential Statistics
- Measures of Central Tendency and Measures of Dispersion
- Boxplots
- Outliers
- Histograms
- Correlation

❸ PARSE the Data

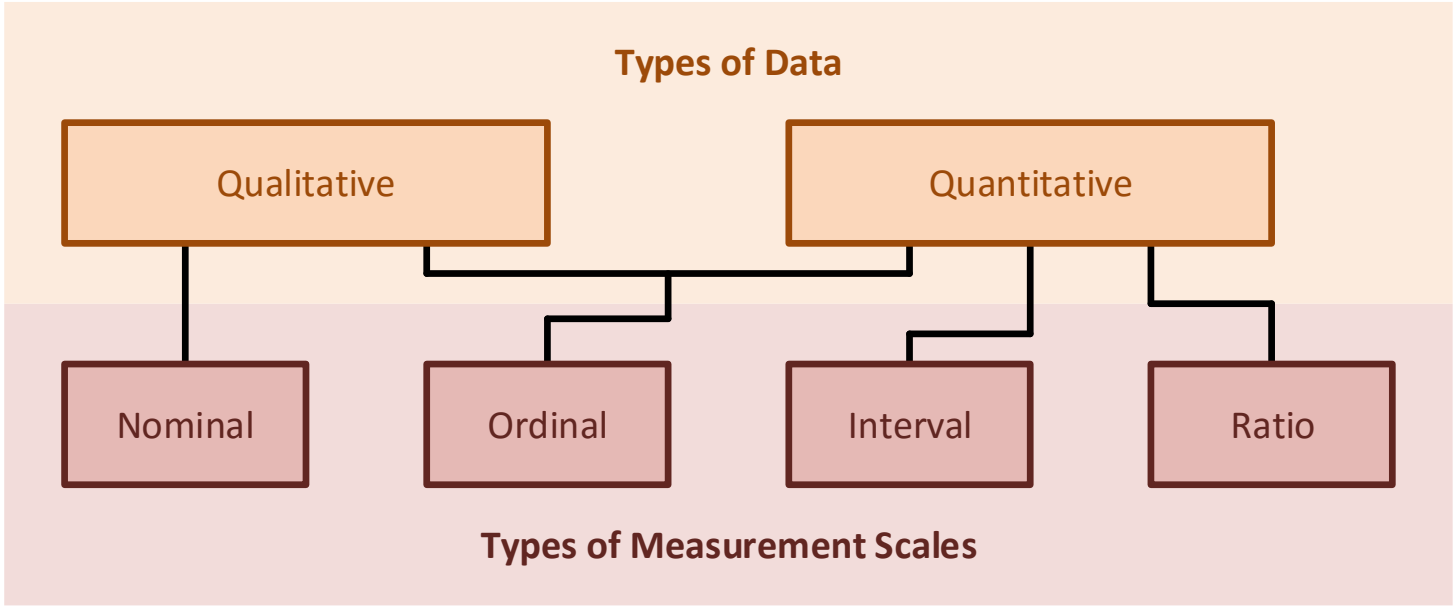
Types of Data and
Types of Measurement Scales

Types of Data



- Qualitative Data
  - Uses descriptive terms to measure or classify something of interest, e.g., education level
- Quantitative Data
  - Uses numerical values to describe something of interest, e.g., age

Types of Measurement Scales



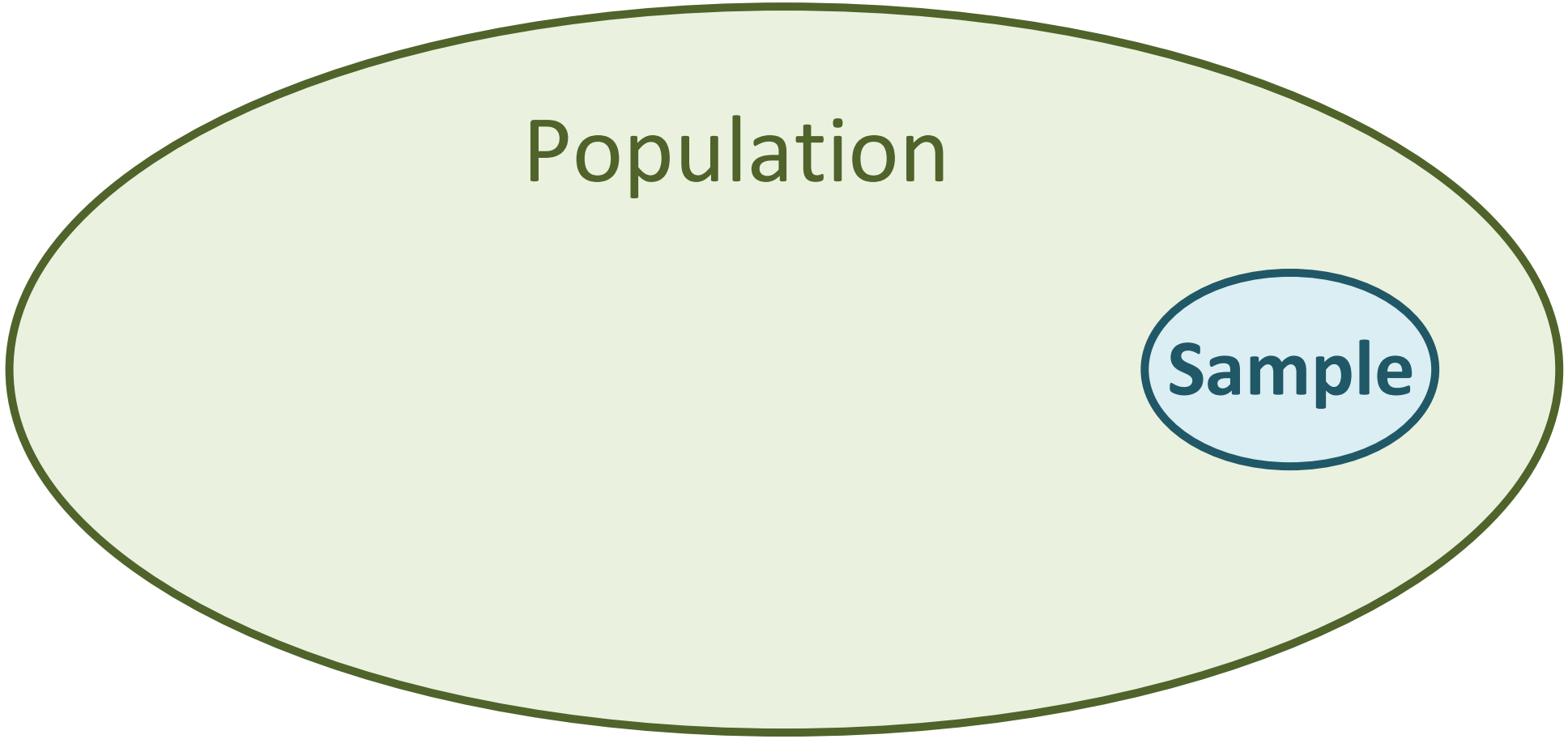
Types of Measurement Scales (cont.)

	Nominal	Ordinal	Interval	Ratio
e.g.	Gender	Movie ratings	Temperature	Salary
Categorize?	✓ (male, female)	✓	✓	✓
Rank-order?	✘	✓ (★<2★<3★<4★)	✓	✓
Add and subtract?	✘	✘ (4★−3★≠★)	✓ (75°C is 50°C warmer than 25°C)	✓
Multiply and divide?	✘	✘ (4★ not 4× better than 1★)	✘ (75°C not 3× as warm as 25°C) (0°C doesn't mean no temperature!)	✓ (Salary of $200K is 2× that of $100K) ($0 means no salary ☹)

❸ PARSE the Data

Populations and Samples

Populations and Samples

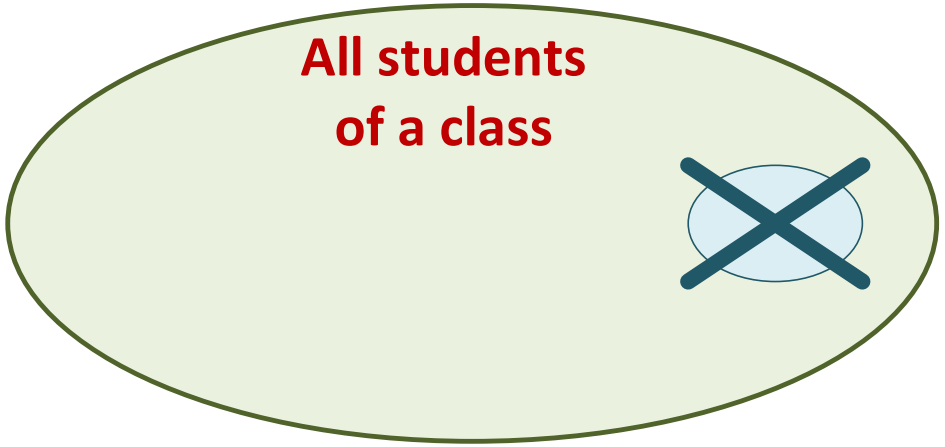


A dataset may be considered either as a population or a sample, depending on the reason for its collection and analysis

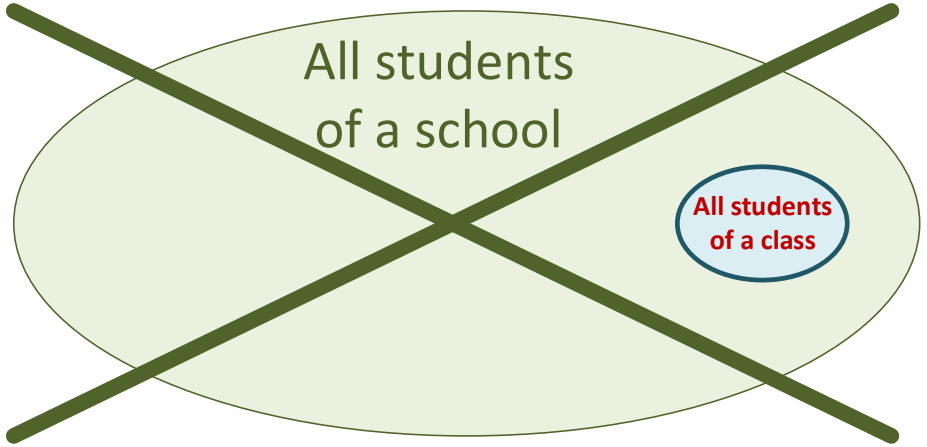
- Students of a class are a population if the analysis describes the distribution of scores in that class

- But they are a sample the analysis infers from their scores the scores of other students (e.g., all students from that school)

Descriptive Statistics



Inferential Statistics

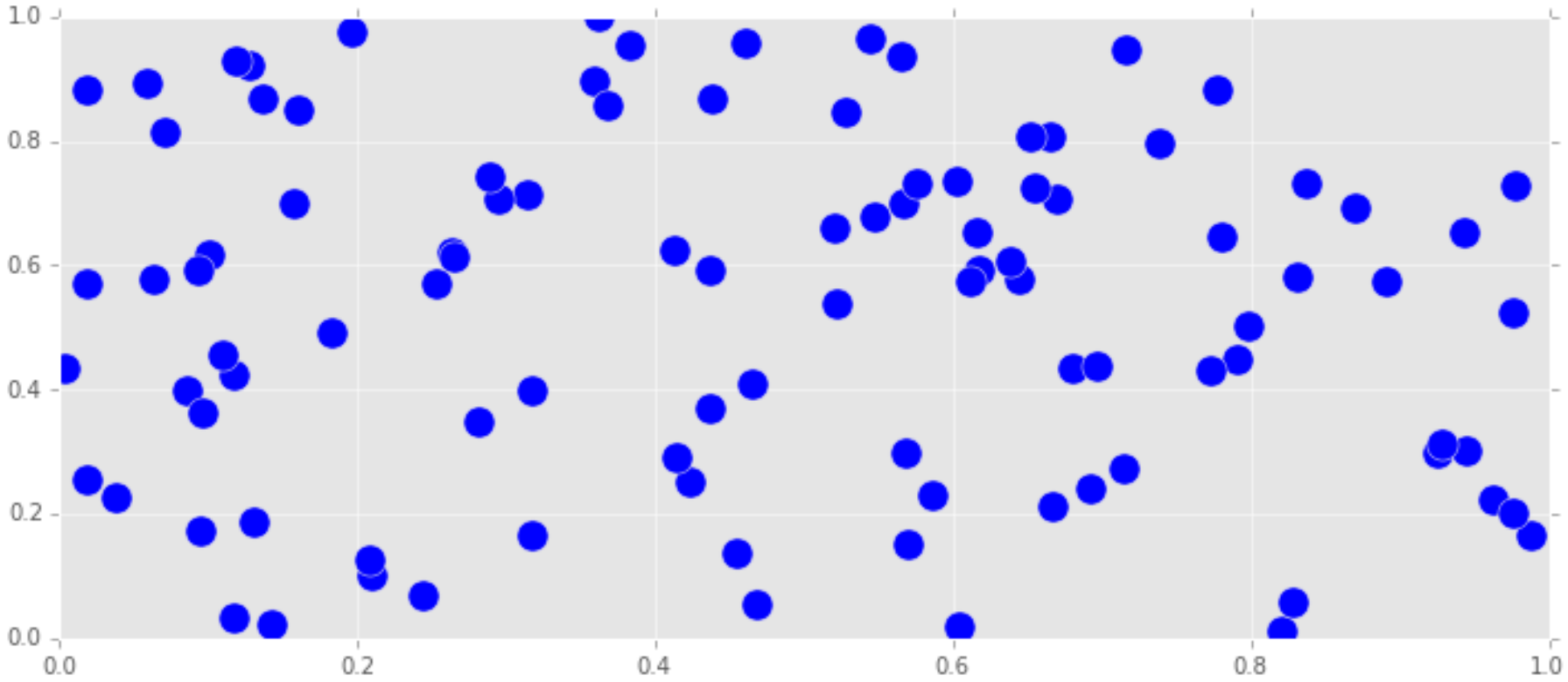


❸ PARSE the Data

Activity | Summarizing Data

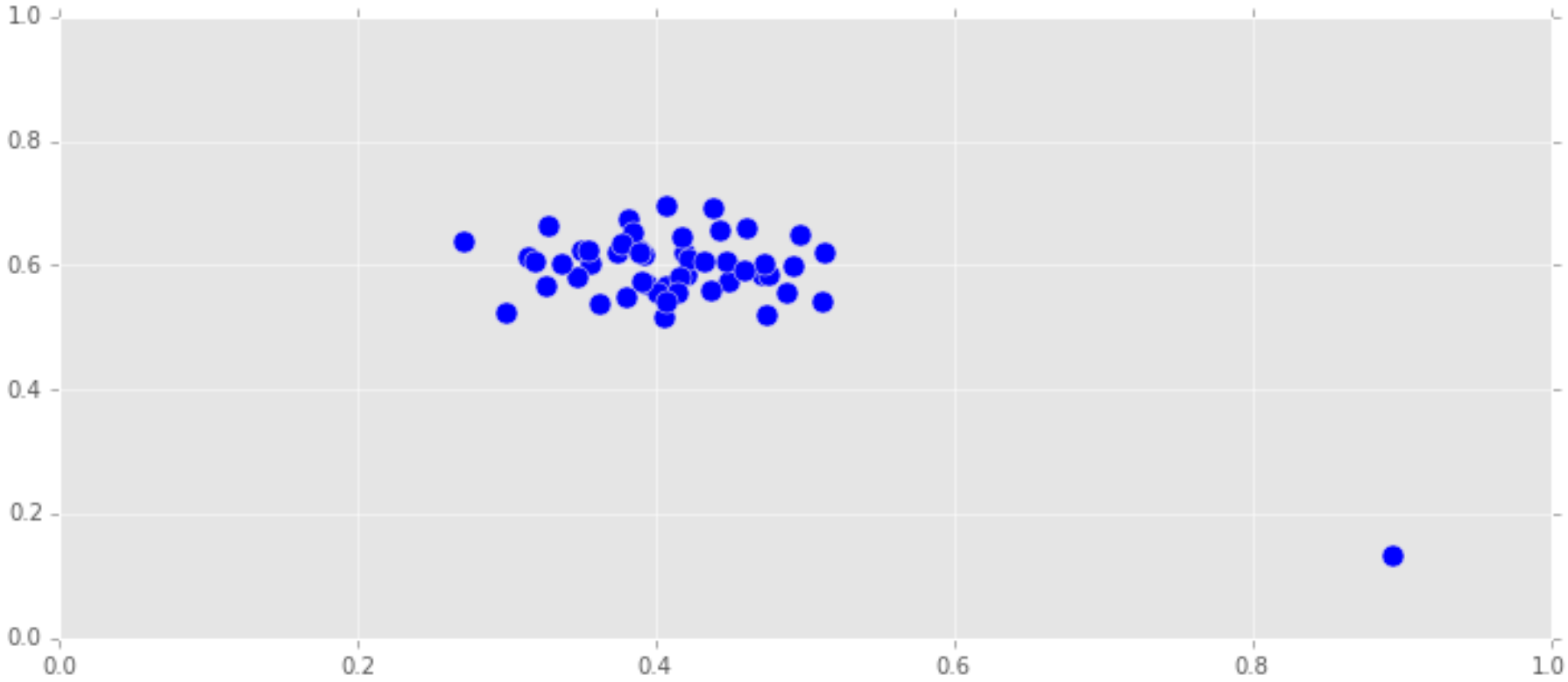
Activity | How would you summarize this data?

EXERCISE



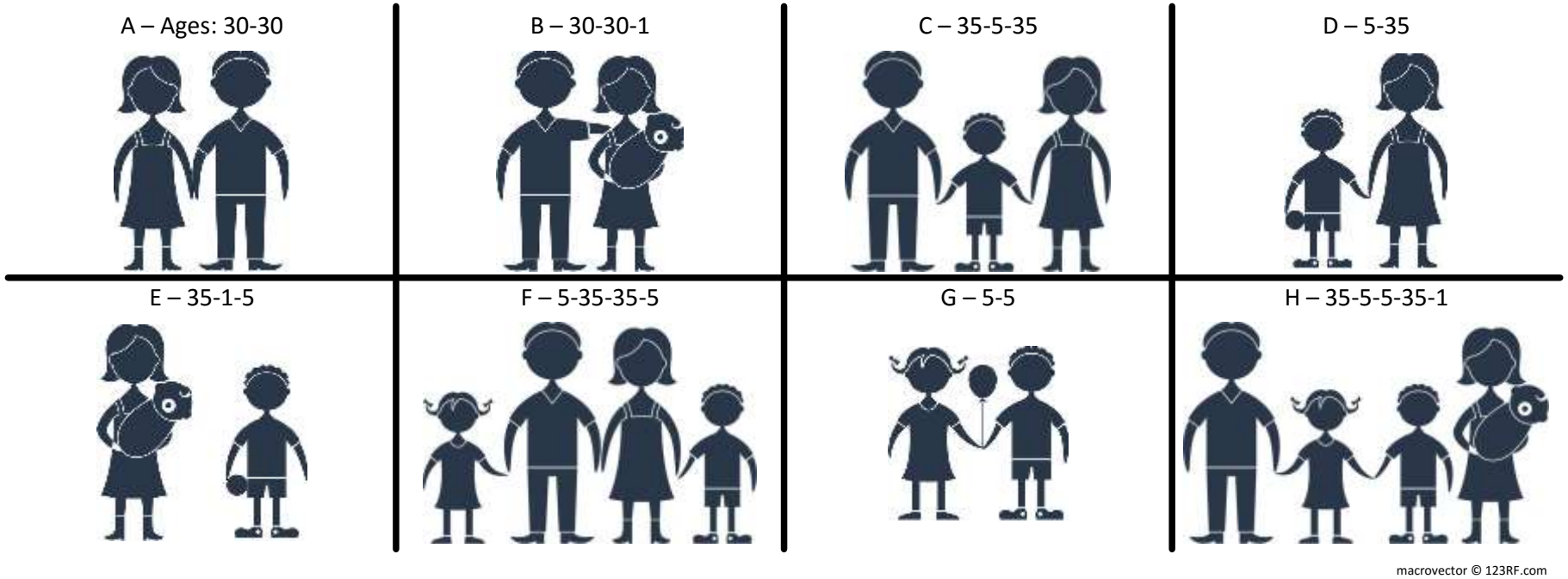
Activity | How would you summarize this data? (cont.)

EXERCISE



Activity | Measures of Central Tendency. What is the typical age for each of these 8 groups of people? (10 minutes)

EXERCISE



Activity | What is the typical age for each of these 8 groups of people? (cont.)

Group	Mean	Median	Mode
A (30-30)	30[1]	30[1]	30[1]
B (30-30-1)	20.3[2] (i.e., no 20-year-olds in the group)	30[3]	30[3]
C (35-5-35)	25[2]	35[3]	35[3]
D (5-35)	20[2]	20[2]	None[4]
E (35-1-5)	13.6[2]	5[2]	None[4]
F (5-35-35-5)	20[2]	20[2]	5 and 35[5]
G (5-5)	5[1]	5[1]	5[1]
H (35-5-5-35-1)	16.2[2]	5[6]	5 and 35[5]

[1] All values are equal [2] Value not representative [3] Follow the "majority" [4] All values are different [5] Follow the "majorities" [6] Partially correct

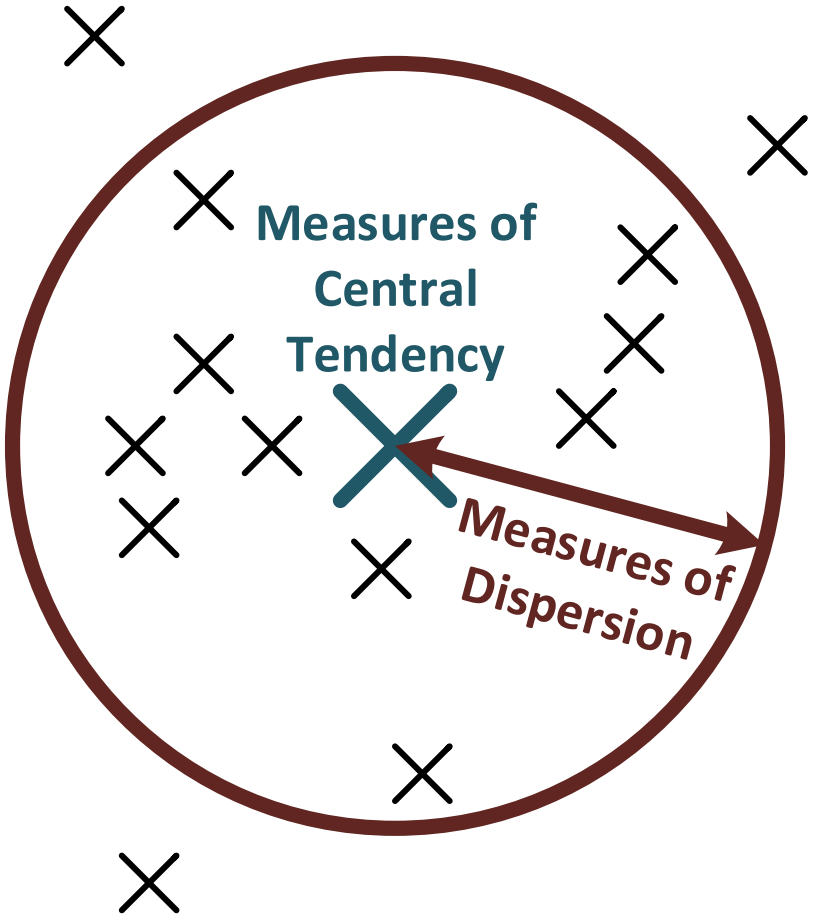
Mean, Median, and Mode: There is no "Winner-Take-All"

	Value is in the dataset	Value is easy to compute	Value is resistant to outliers	Corresponding measure of Dispersion	Used extensively by mathematical models
Mean	☹ (Unlikely)	☺	☹	☺ (Variance, Standard Deviation)	☺
Median	😐 (50% chance)	😐 (need to rank the values)	☺	☺ (Interquartile Range)	☹
Mode	☺ (Always)	☹ (Need to count and rank the count)	☺	☹ (Not really)	☹ (Mode might not be defined or you might have multiple values)

❸ PARSE the Data

Measures of Central Tendency and Measures of Dispersion

Measures of Central Tendency and Measures of Dispersion



- Measures of Central Tendency
  - (Or measures of location)
  - Answer the question: “What’s the typical or common value for a variable?”
  - Mean, Median, Mode
- Measures of Dispersion
  - (Or measures of variability/spread)
  - Answer the question: “How far do values stray from the typical value?”
  - Variance, Standard Deviation, Range, Interquartile Range (IQR)

(Arithmetic) Mean, Variance, and Standard Deviation

Ordinal ✘	Nominal ✘	Interval ✓	Ratio ✓

	Population	Sample
(Arithmetic) Mean *(a.k.a., the first moment)* (Mean has unit of X:$[X]$)	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = E[X^1]$ (mu)	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (x-bar)
Variance *(a.k.a., the second moment)* $[X^2]$	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ $= E[(X-\mu)^2]$ (sigma-squared)	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard Deviation $[X]$	$\sigma = \sqrt{\sigma^2}$ (sigma)	$s = \sqrt{s^2}$

(mean, variance, and standard deviations are based on the values of x_i)

❸ PARSE the Data

Codealong – Part A

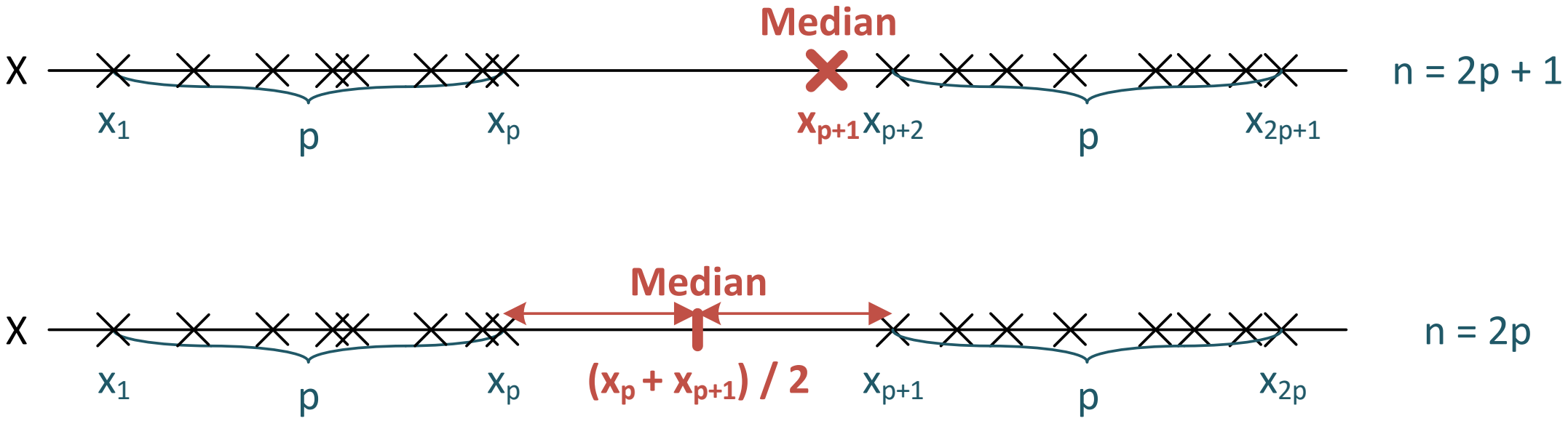
`.mean()`

`.var(), .std()`

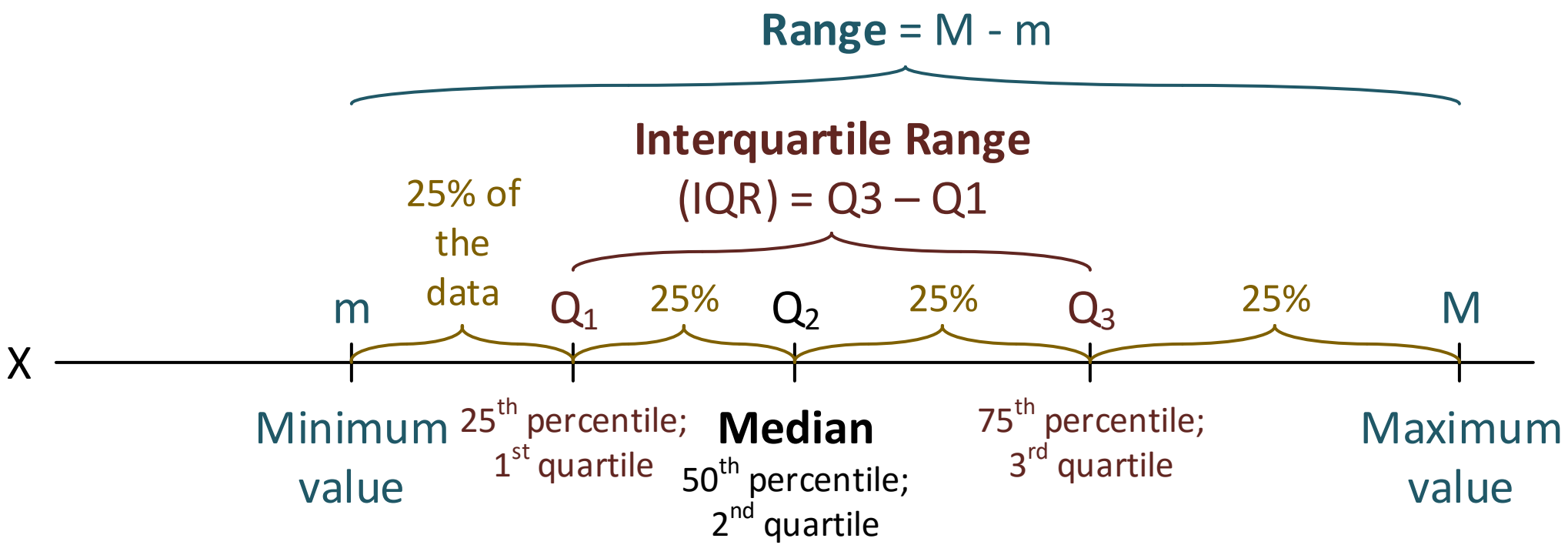
❸ PARSE the Data

Median, Range, and Interquartile Range

Median


Median
x
n = 2p + 1
x1
p
xp
xp+1 xp+2
p
x2p+1
Median
x
n = 2p
x1
p
xp
(xp + xp+1) / 2
xp+1
p
x2p

Median, Range, and Interquartile Range



Median, Range, and Interquartile Range (cont.)

Nominal ✘	Ordinal ✘	Interval ✓	Ratio ✓

Median	$median = \begin{cases} x_{p+1} \; if \; n = 2p+1 \\ \dfrac{x_p + x_{p+1}}{2} \; if \; n = 2p \end{cases}$
Range	$range = x_n - x_1$
Percentile	$q_k = \begin{cases} x_{\lceil p \rceil} \; if \; p = \dfrac{nk}{100} \; not \; integer \\ \dfrac{x_p + x_{p+1}}{2} \; otherwise \end{cases}$
Quartile	$Q_1 = q_{25}; Q_3 = q_{75}$
Interquartile Range	$IQR = Q_3 - Q_1$

(median, range, and interquartile range are based on the ranks of x_i; x_i ranked from smallest to largest)

❸ PARSE the Data

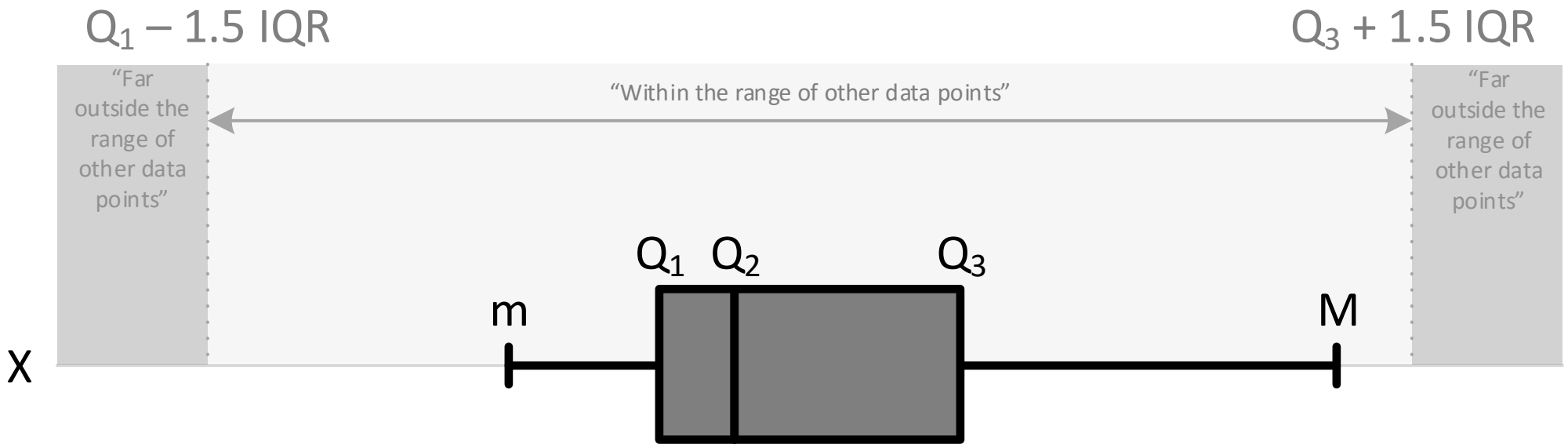
Codealong – Part B

.mean(), .median()
.count(), .dropna(), .isnull()
.min(), .max()
.quantile()
.describe()

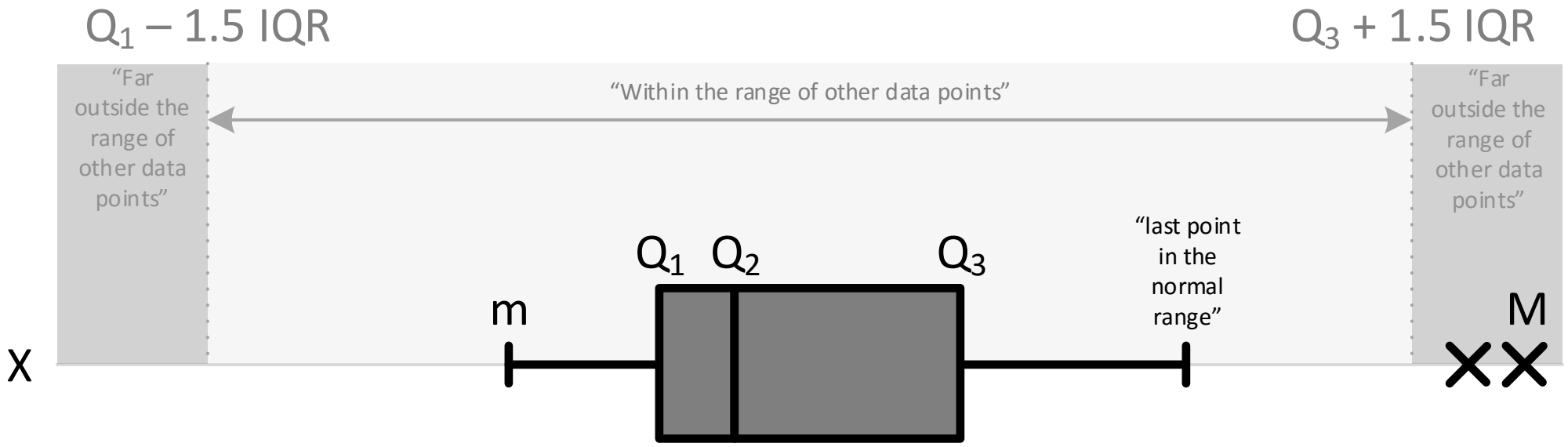
❸ PARSE the Data

Median, Range, Interquartile Range, and Boxplots

Boxplot #1 | Median, Range, Interquartile Range, and no Outliers



Boxplot #2 | Median, Range, Interquartile Range, and Outliers



❸ PARSE the Data

Codealong – Part C

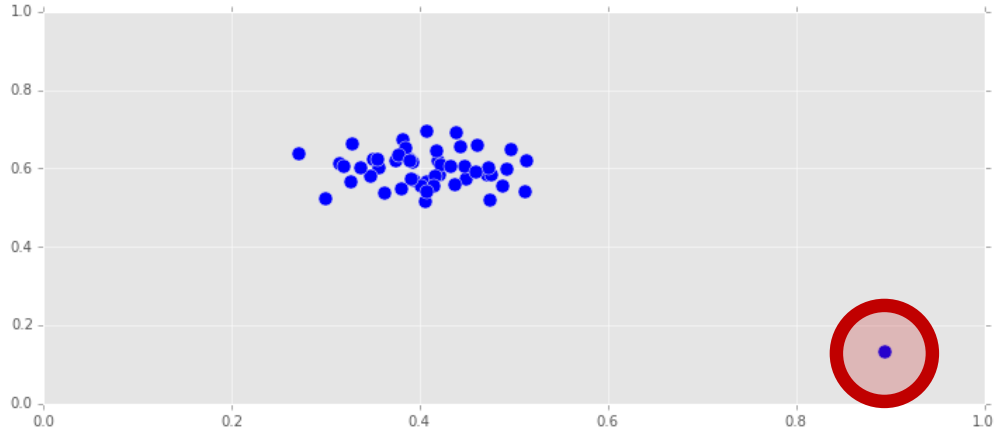
Boxplots

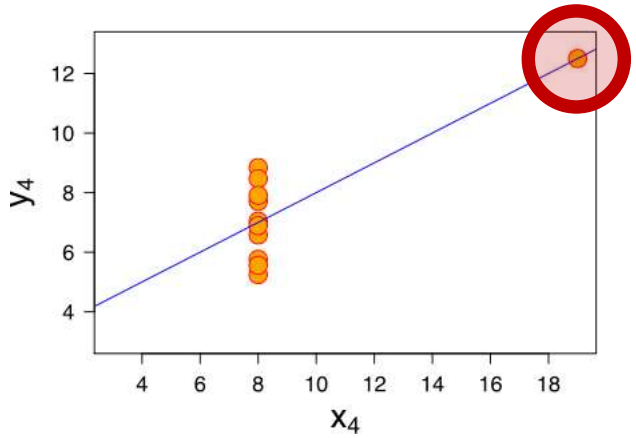
❸ PARSE the Data

Outliers

Think twice before discarding outliers; they might be the most important points

- Outliers are values that are “far” from the central tendency
- No formal definition among statisticians on how to define outliers (how do you define “far”?)
- However, general agreement that they be identified and dealt with appropriately (e.g., keep or discard)
  - They might be the most important points of your dataset

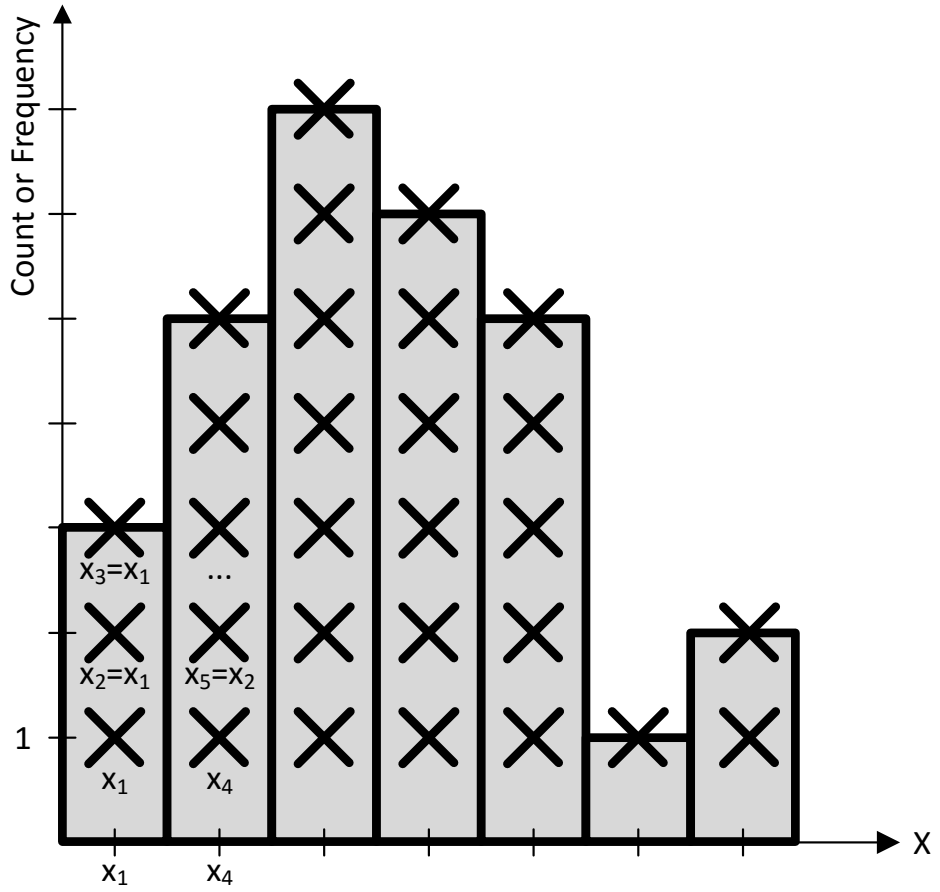




❸ PARSE the Data

Histograms

Histograms. $x_1 = x_2 = x_3 < x_4 = x_5 \ldots$



❸ PARSE the Data

Codealong – Part D

Histograms

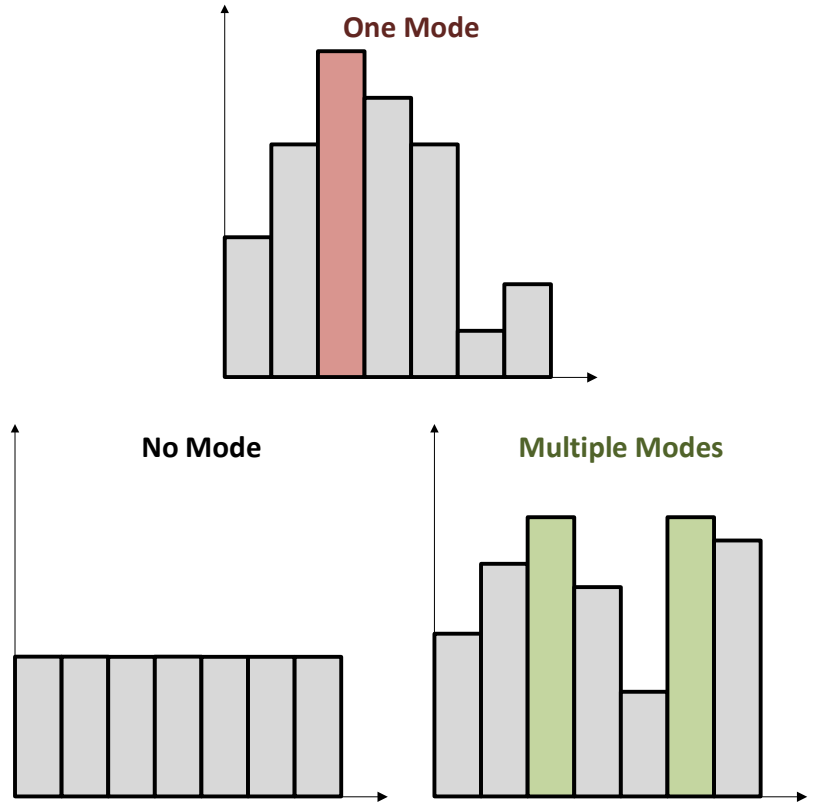
❸ PARSE the Data

Mode

Modes and Histograms

Nominal ✓	Ordinal ✓	Interval ✓	Ratio ✓

- The Mode is the value(s) that occur(s) most often



❸ PARSE the Data

Codealong – Part E

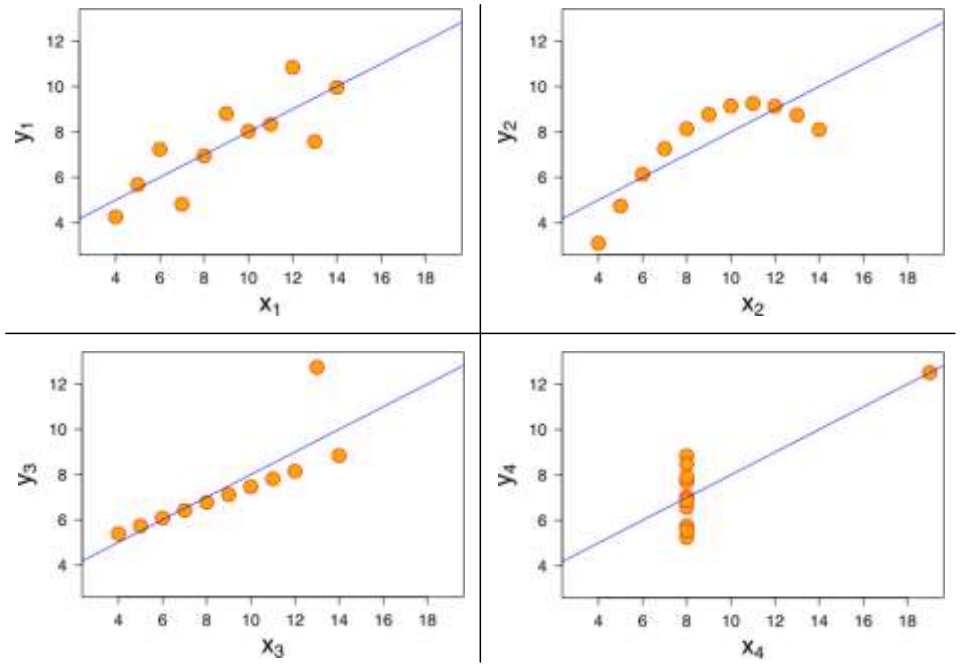
`.mode()`

❸ PARSE the Data

Plot the Data!

Don’t rely on basic statistic properties and **plot the data!** 4 datasets (Anscombe’s quartet) that have nearly identical simple statistical properties, yet are very different

Scatter plot appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.



Not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the linear correlation is not relevant.

Distribution is linear, but with a different regression line, which is offset by the one outlier which exerts enough influence to alter the regression line.

Example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

Property	Value
Mean of x_i	9
Sample variance of x_i	11
Mean of y_i	7.50
Sample variance of y_i	4.122 or 4.127
Correlation between x_i and y_i	0.816
Linear regression line in each case	$y_i = 3.00 + 0.500\ x_i$

❸ PARSE the Data

(Linear) Correlation

Correlation

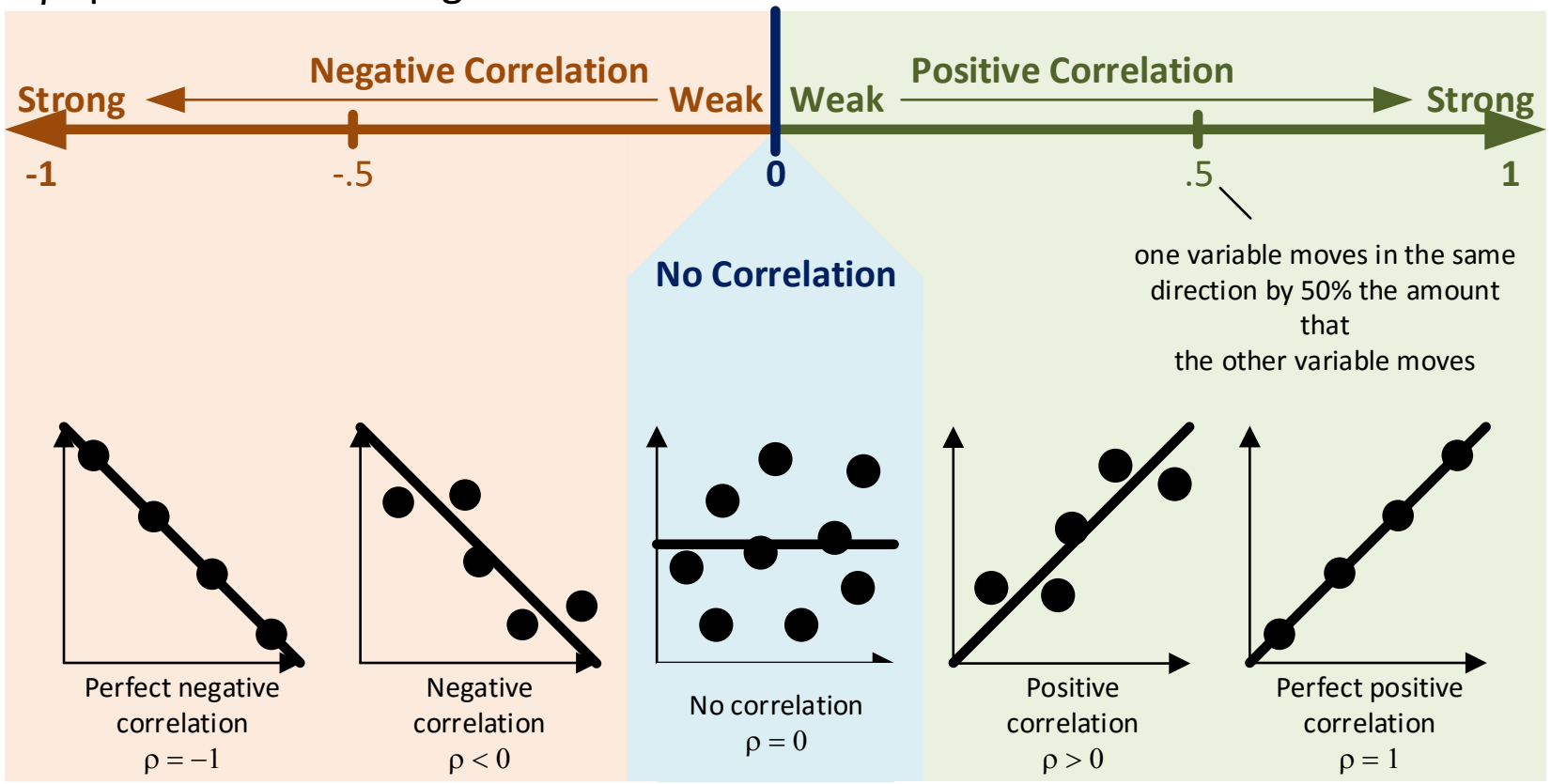
- A measure of strength and direction for a **linear association** between two random variables

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

- $\rho = 0$ means that the two variables don't have a linear association
  - It doesn't imply that they are independent!

Correlation (cont.)

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

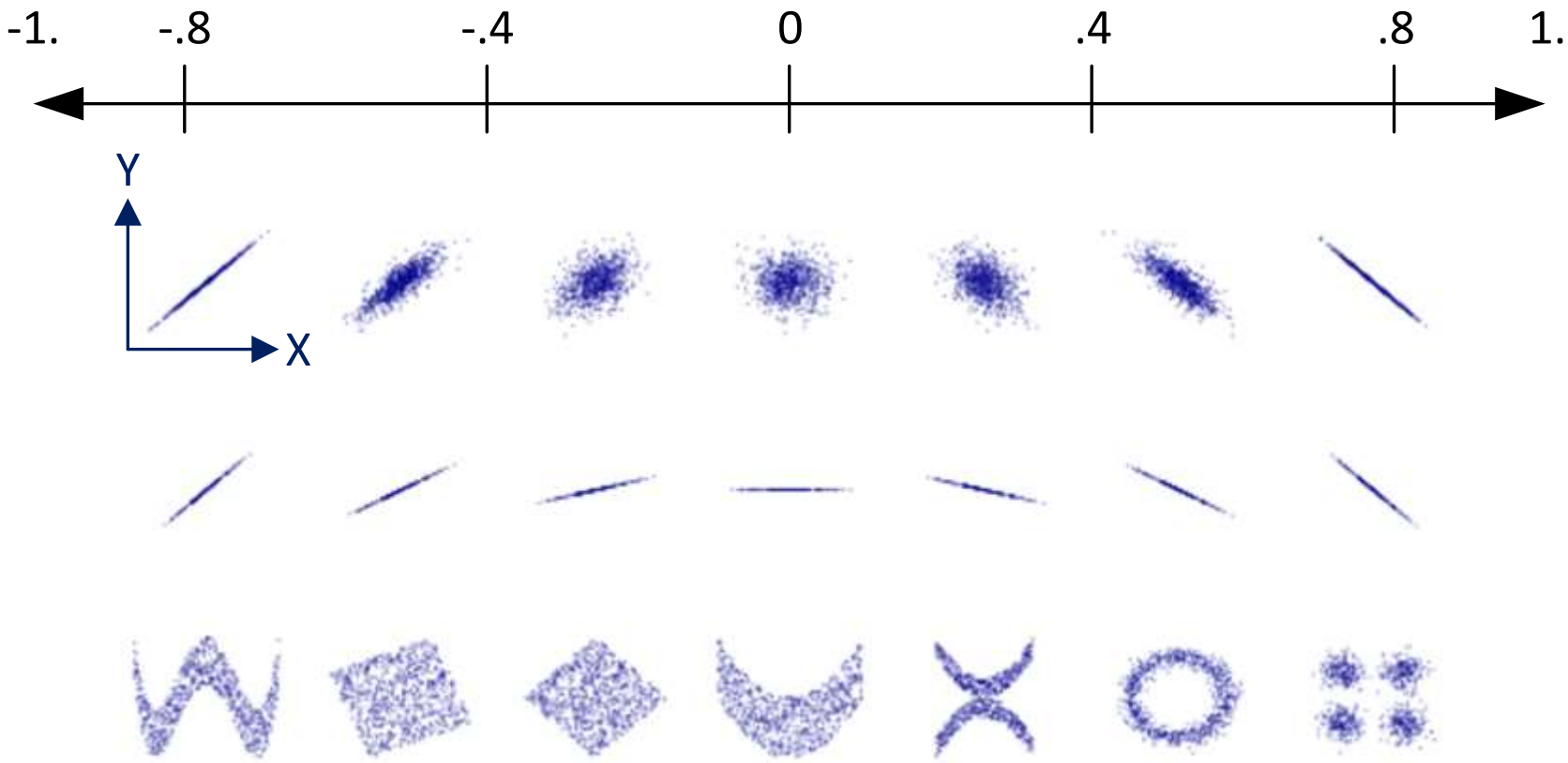


❸ PARSE the Data

Activity | Correlations and Scatter Plots

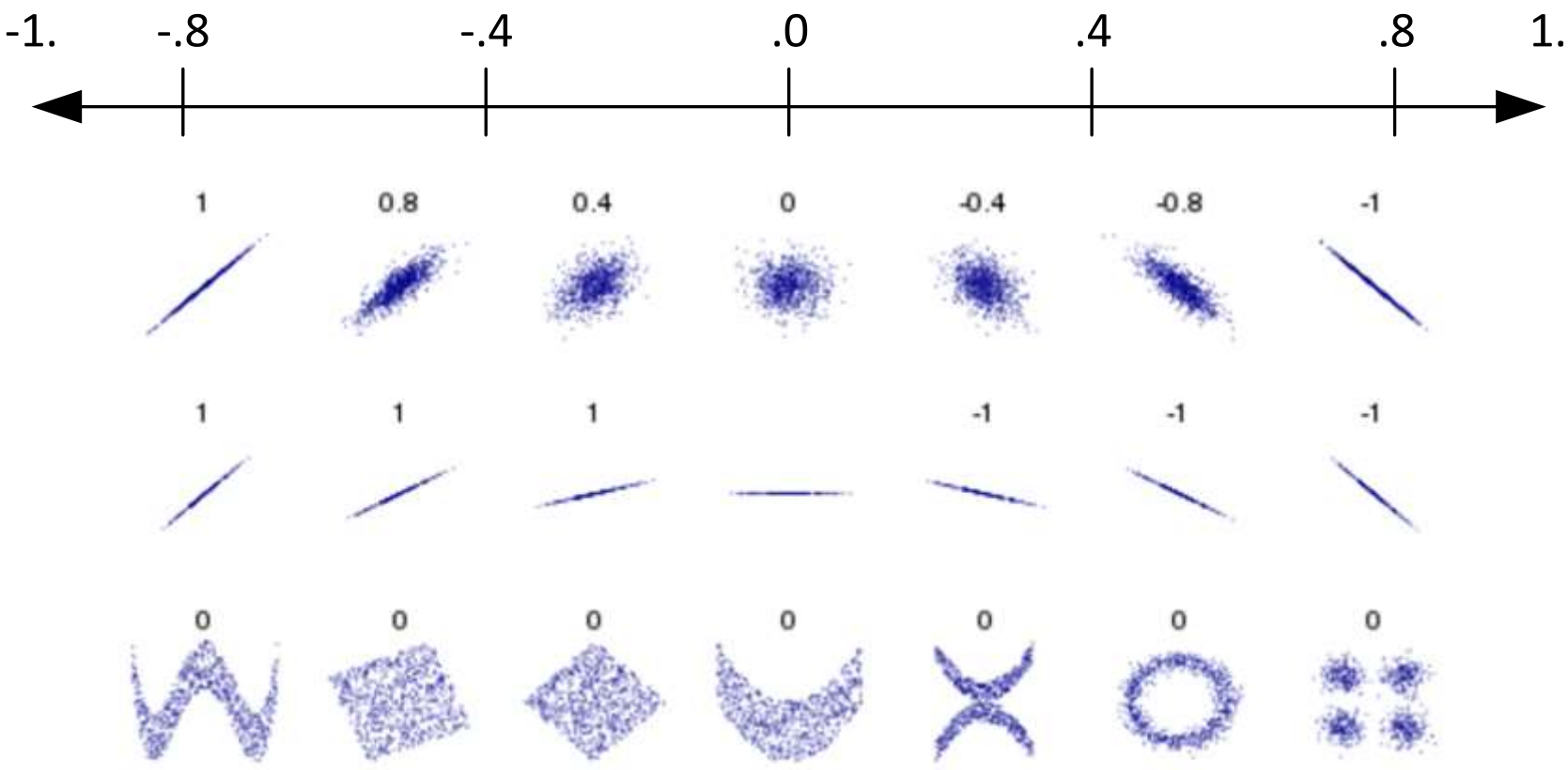
Activity | What's the correlations for the following scatter plots (5 minutes)

EXERCISE



Activity | What's the correlations for the following scatter plots (cont.)

EXERCISE



Source: Wikipedia

❸ PARSE the Data

Codealong – Part F

`.corr()`

Heatmaps
Scatter plots and matrices

❸ PARSE the Data

Codealong – Part G

`.value_counts()`

`.crosstab()`

Lab

Exploratory Data Analysis with pandas

Review

Review

You should now be able to:

- Identify variable types
- Use the *pandas* (and *NumPy*) libraries to analyze datasets using basic summary statistics: mean, median, mode, max, min, quartile, inter-quartile range, variance, standard deviation, and correlation
- Create data visualizations – including boxplots, histograms, and scatter plots – to discern characteristics and trends in a dataset

Q & A

Next Class

Flexible Class Session #1 | Exploratory Data Analysis

Exit Ticket

Don’t forget to fill out your exit ticket here