# 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

DS

# Announcements and Exit Tickets

DS

# Review

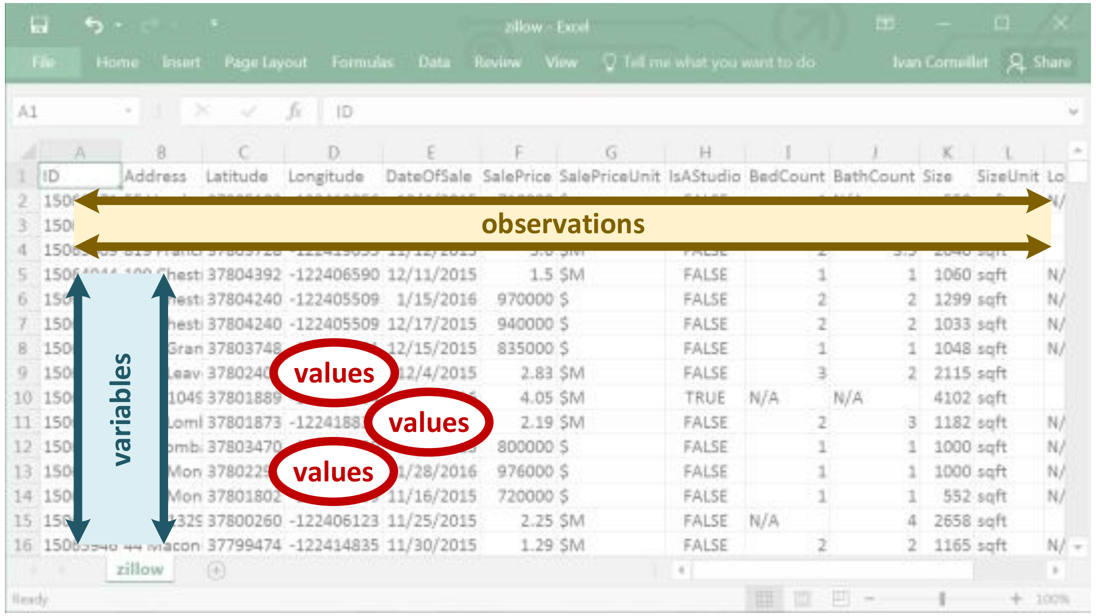
# Review

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



|  | A | B | C | D | E | F | G | H | I | J | K | L |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ID | Address | Latitude | Longitude | DateOfSale | SalePrice | SalePriceUnit | IsAStudio | BedCount | BathCount | Size | SizeUnit | Lo |
| 2 | 150 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | N/ |
| 3 | 150 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |  |
| 4 | 150 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |  |
| 5 | 150 | Chest | 37804392 | -122406590 | 12/11/2015 | 1.5 | $M | FALSE | 1 | 1 | 1060 | sqft | N/ |
| 6 | 150 | hest | 37804240 | -122405509 | 1/15/2016 | 970000 | $ | FALSE | 2 | 2 | 1299 | sqft | N/ |
| 7 | 150 | hest | 37804240 | -122405509 | 12/17/2015 | 940000 | $ | FALSE | 2 | 2 | 1033 | sqft | N/ |
| 8 | 150 | Gran | 37803748 | [illegible] | 12/15/2015 | 835000 | $ | FALSE | 1 | 1 | 1048 | sqft | N/ |
| 9 | 150 | Leav | 3780240 | [illegible] | 12/4/2015 | 2.83 | $M | FALSE | 3 | 2 | 2115 | sqft |  |
| 10 | 150 | 1049 | 37801889 | [illegible] | [illegible] | 4.05 | $M | TRUE | N/A | N/A | 4102 | sqft |  |
| 11 | 150 | Lomi | 37801873 | -1224188 | [illegible] | 2.19 | $M | FALSE | 2 | 3 | 1182 | sqft | N/ |
| 12 | 150 | omb | 37803470 | [illegible] | [illegible] | 800000 | $ | FALSE | 1 | 1 | 1000 | sqft | N/ |
| 13 | 150 | Mon | 3780229 | [illegible] | 1/28/2016 | 976000 | $ | FALSE | 1 | 1 | 1000 | sqft | N/ |
| 14 | 150 | Mon | 37801802 | [illegible] | 11/16/2015 | 720000 | $ | FALSE | 1 | 1 | 552 | sqft | N/ |
| 15 | 150 | 1329 | 37800260 | -122406123 | 11/25/2015 | 2.25 | $M | FALSE | N/A |  | 4 | 2658 | sqft |  |
| 16 | 150 | Macon | 37799474 | -122414835 | 11/30/2015 | 1.29 | $M | FALSE | 2 | 2 | 1165 | sqft | N/ |

# Review and Activity | Subsetting with *pandas*

EXERCISE

## DIRECTIONS (10 minutes)

1. Using the dataset in the codealong (Part A), answer the following questions:
    1. Subset the dataframe on the age and gender columns.
    2. Subset the dataframe on the age column alone, first as a *DataFrame*, then as a *Series*
    3. Subset the dataframe on the rows Bob and Carol
    4. Subset the dataframe on the row Eve alone, first as a *DataFrame*, then as a *Series*
    5. How old is Frank?

2. Annotate the next handout with the syntax on how to subset dataframes in *pandas*

3. When finished, share your answers with your table

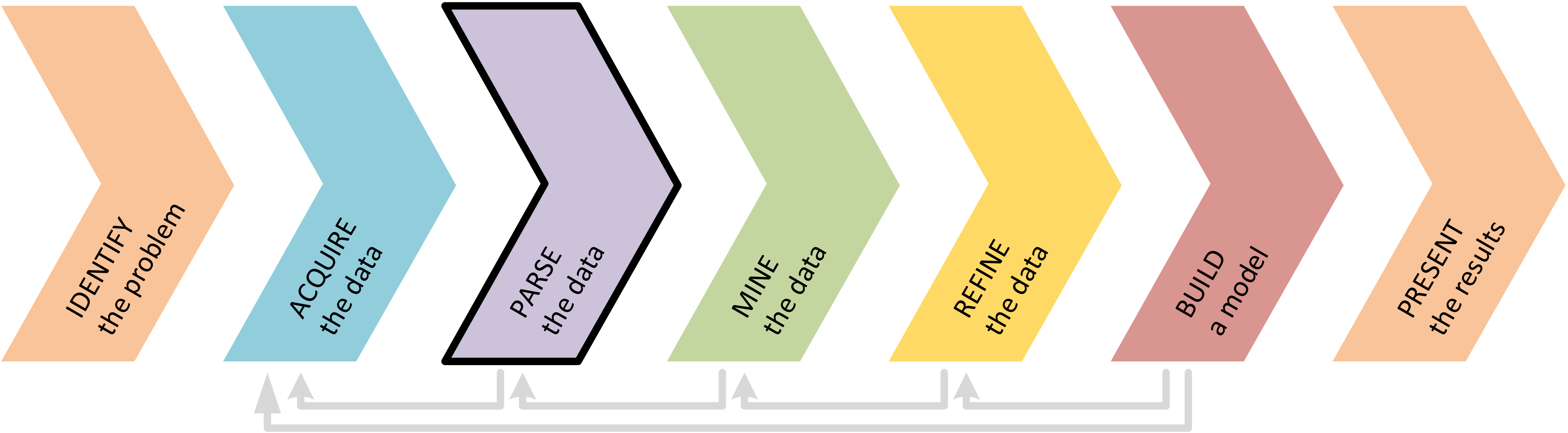
## DELIVERABLE

Answers to the above questions

| | DataFrame | Series |
|---|---|---|
| **Column subsetting** | | |
| **by name**<br>(Columns names are stored in `df.columns`)<br>(`df.columns.get_loc('X1')` returns X1’s column index) | `# New DataFrame with column named X1`<br>`df[ ['X1'] ]`<br><br>`# 2+ columns (in the order listed)`<br>`df[ ['X1', 'X2', …] ]` | `df['X1']`<br><br>`df.X1` |
| **by location** | `# New DataFrame with column at location i (numbering starts at 0)`<br>`df[ [column_i] ]`<br><br>`# 2+ columns (in the order listed)`<br>`df[ [column_i, column_j, …] ]` | |
| **Row subsetting** | | |
| **by index label** | `df.loc[ [index_label_i] ]`<br>`df.loc[ [index_label_i, index_label_j, …] ]`<br><br>`# Can use a range if the index is made of numbers (rows “a” to “b” included)`<br>`df.loc[ index_label_a : index_label_b ]` | `df.loc[index_label_i]` |
| **by location** | `df.iloc[ [row_i] ]`<br>`df.iloc[ [row_i, row_j, …] ]`<br><br>`# (rows “a” to “b’ excluded)`<br>`df.iloc[row_a : row_b ] or df[row_a : row_b ]` | `df.iloc[location_i]` |
| **Cell/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 ❸ **PARSE** the data



IDENTIFY the problem

ACQUIRE the data

PARSE the data

MINE the data

REFINE the data

BUILD a model

PRESENT the results

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

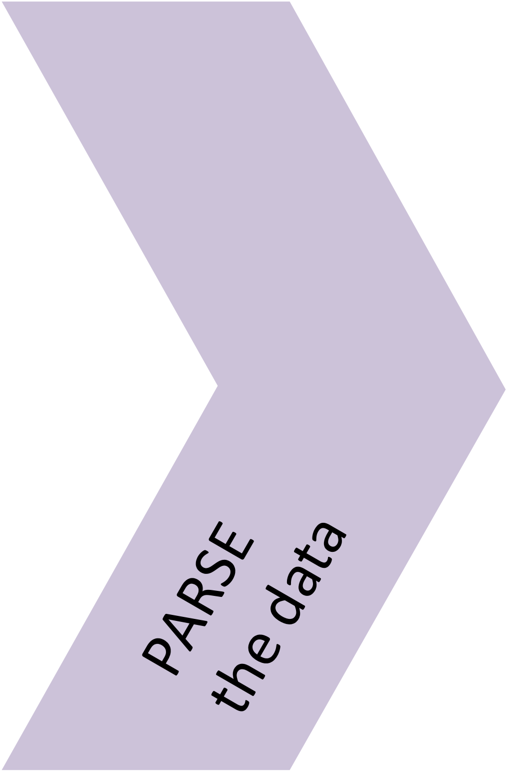
| | | | | |
|---|---|---|---|---|
| **Research Design and Data Analysis** | Research Design | Data Visualization in *pandas* | Statistics | Exploratory Data Analysis in *pandas* |
| **Foundations of Modeling** | Linear Regression | Classification Models | Evaluating Model Fit | Presenting Insights from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

# ❸ PARSE the Data

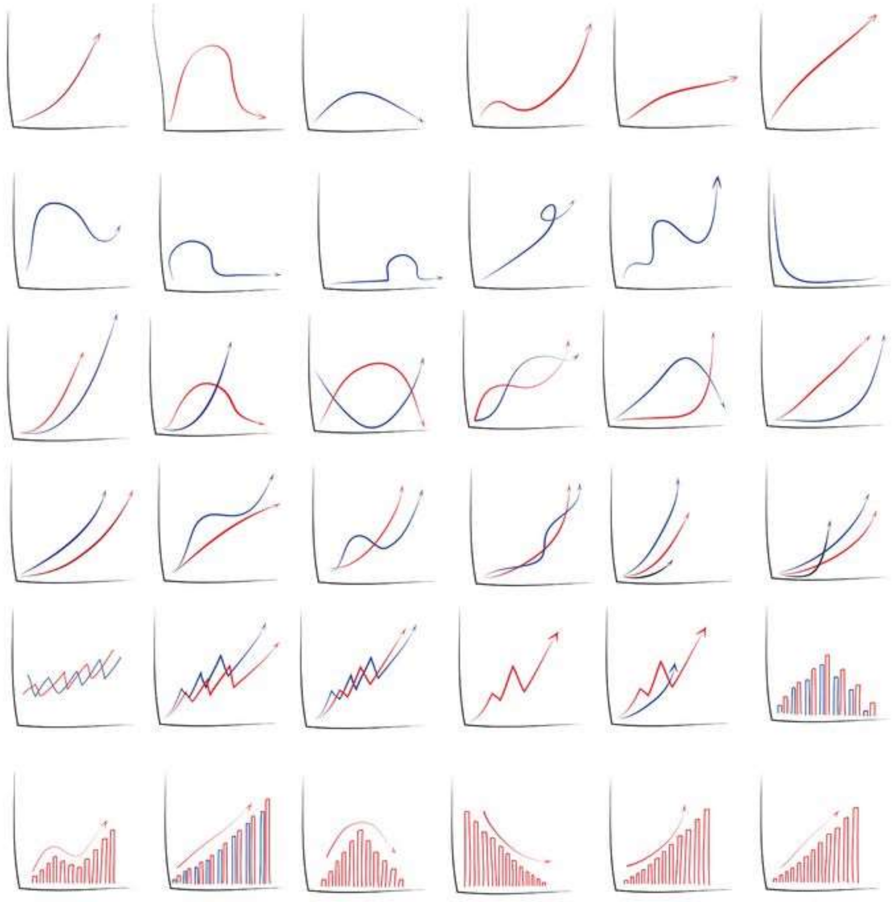
# ❸ Parse the Data



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



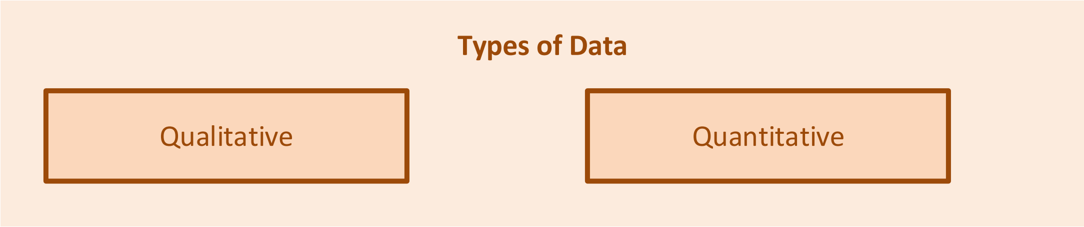
Napat Polchoke © 123RF.com

- 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



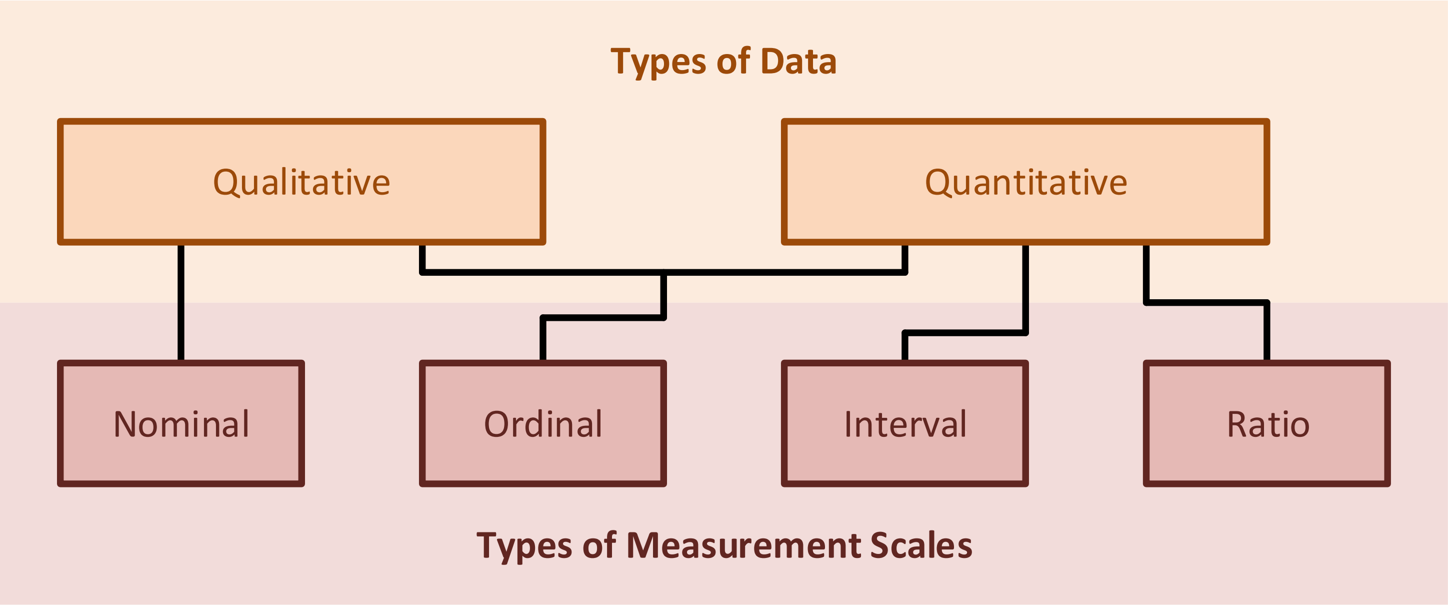
Types of Data

Qualitative

Quantitative

- 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



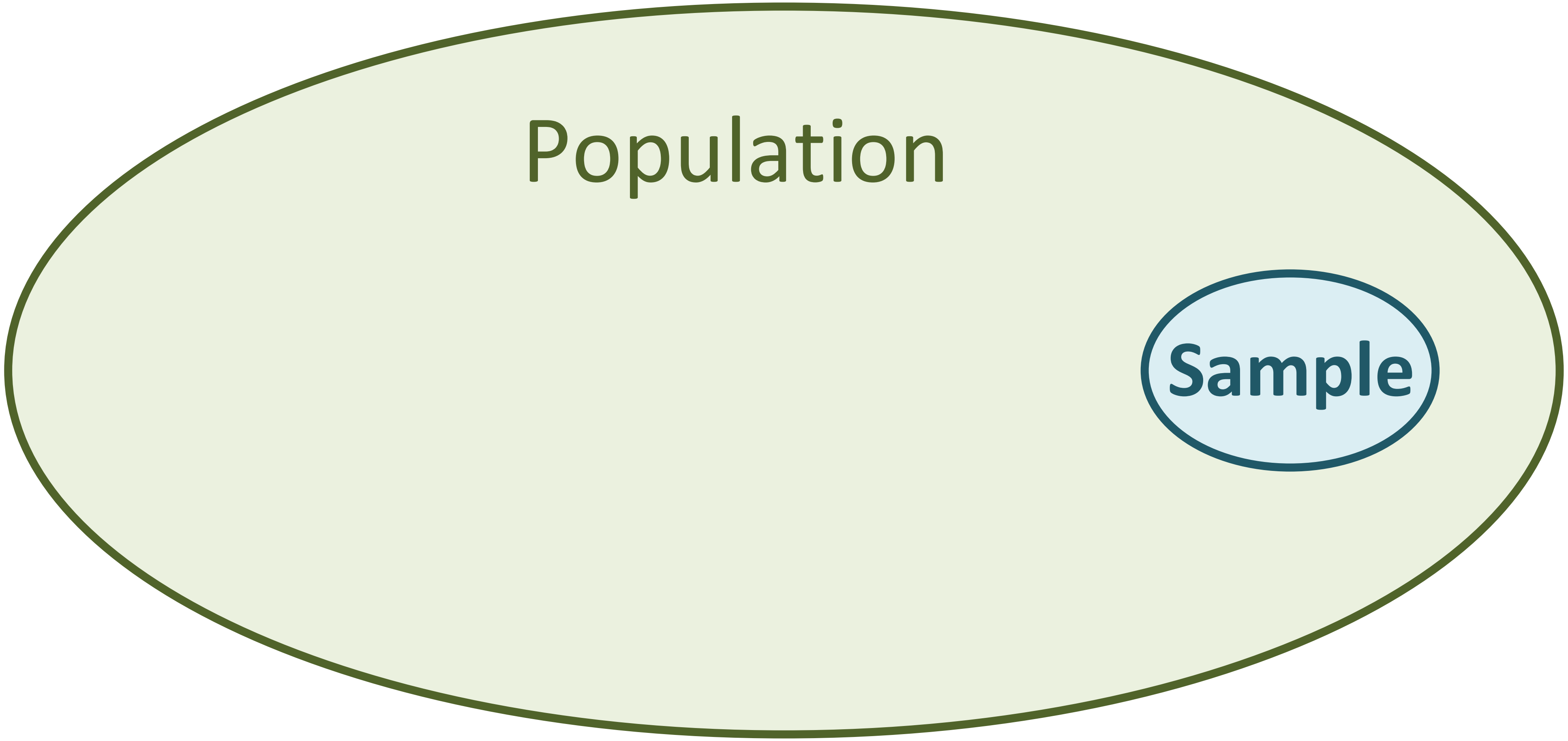
# Types of Measurement Scales (cont.)

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

# ❸ PARSE the Data

*Populations and Samples*

# Populations and Samples

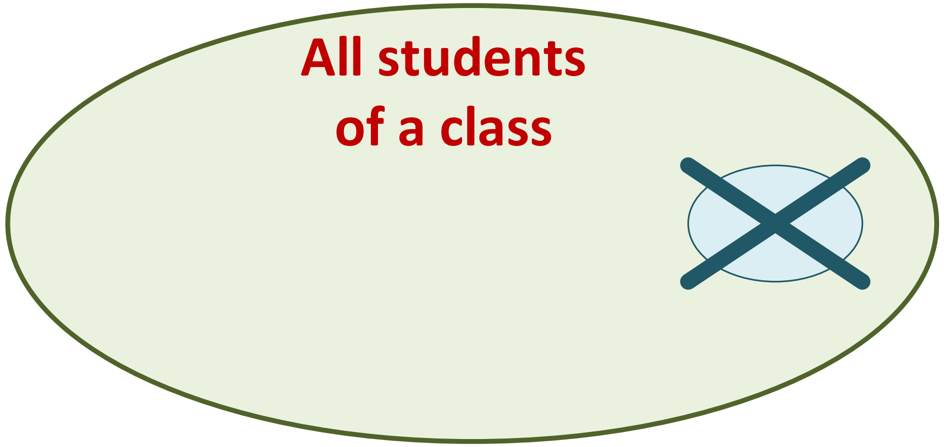


# A dataset may be considered either as a population or a sample, depending on the reason for its collection and analysis

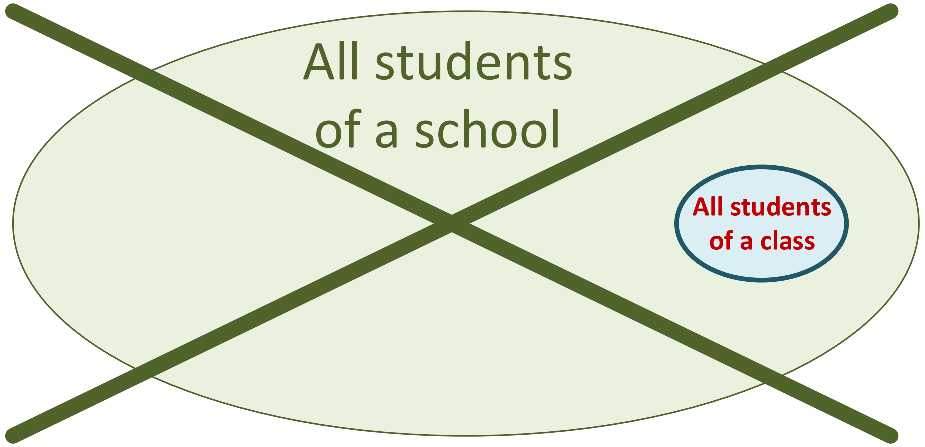
- Students of a class are a population if the analysis describes the distribution of scores in that class

- 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



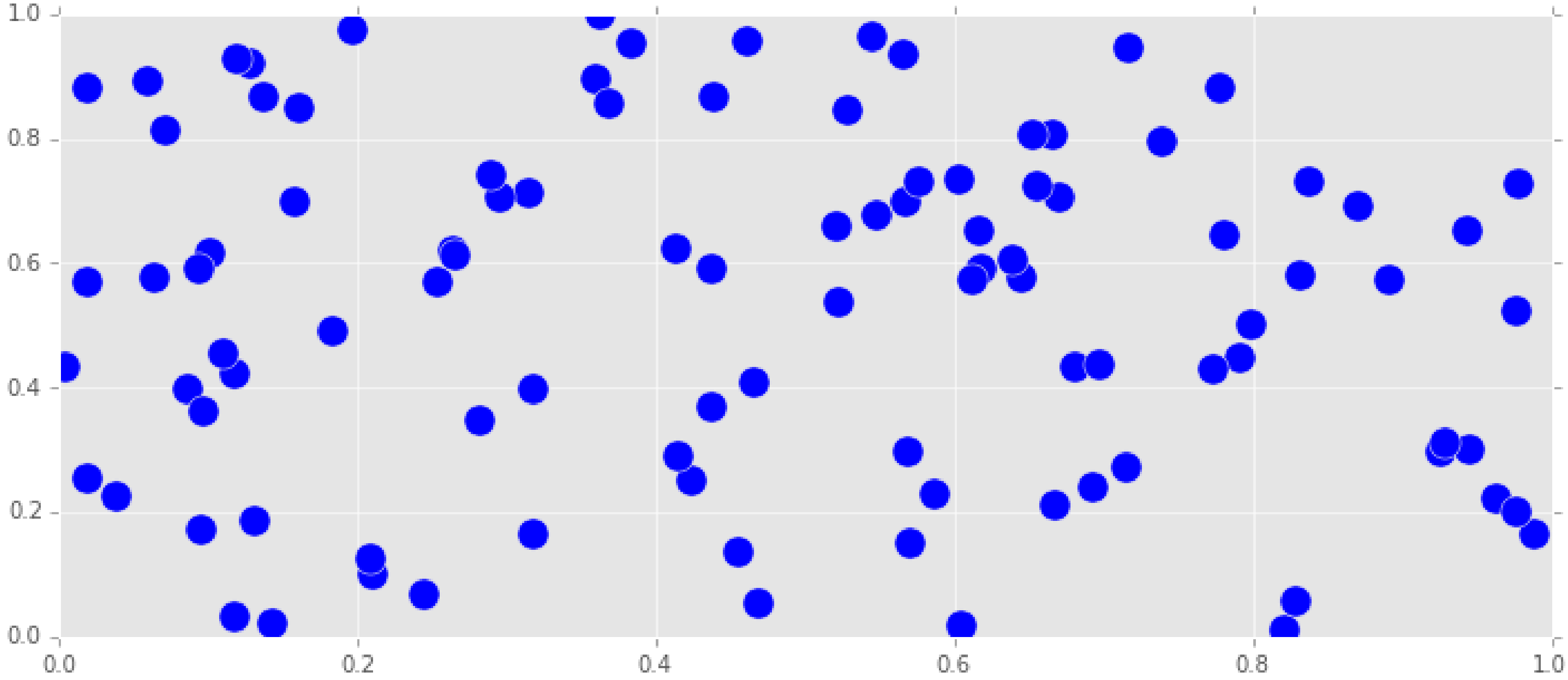
DS

# ❸ PARSE the Data

*Activity | Summarizing Data*

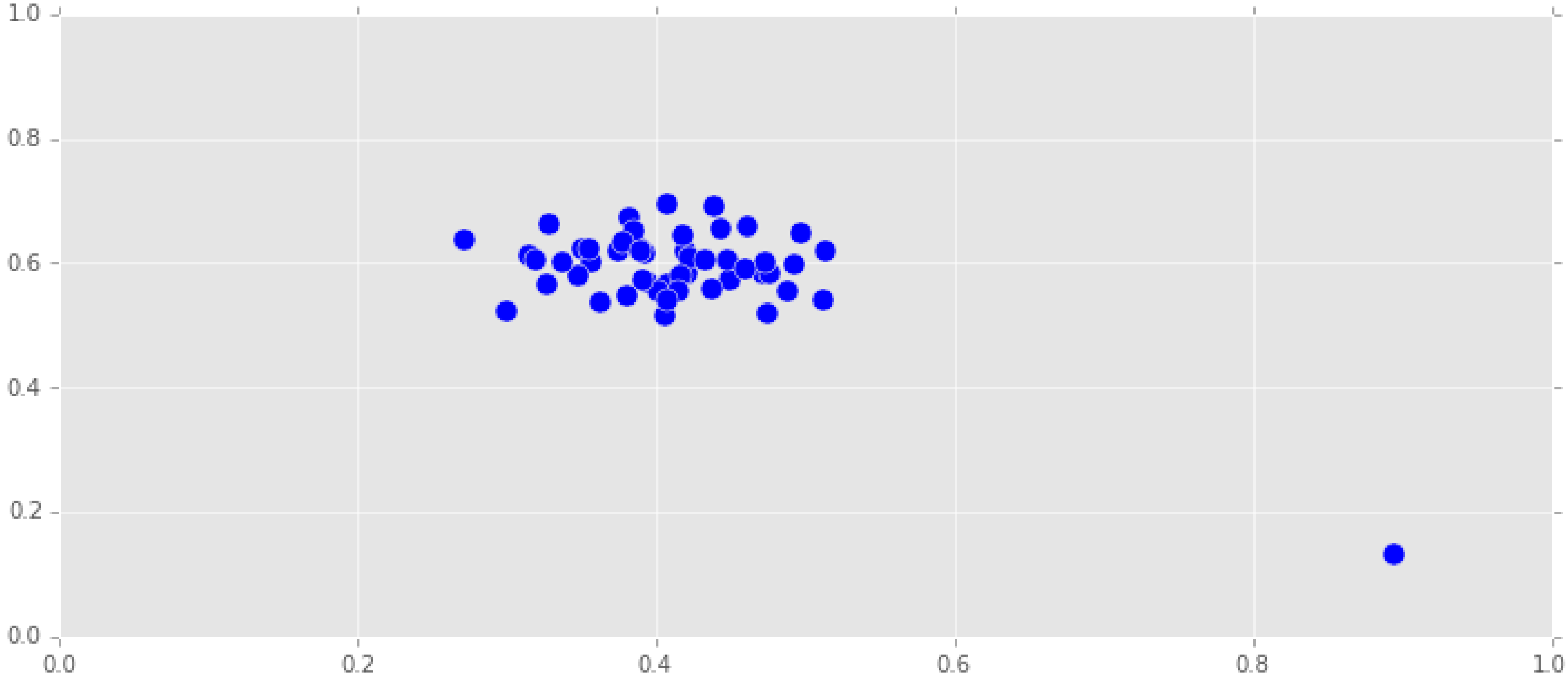
# Activity | How would you summarize this data?

EXERCISE



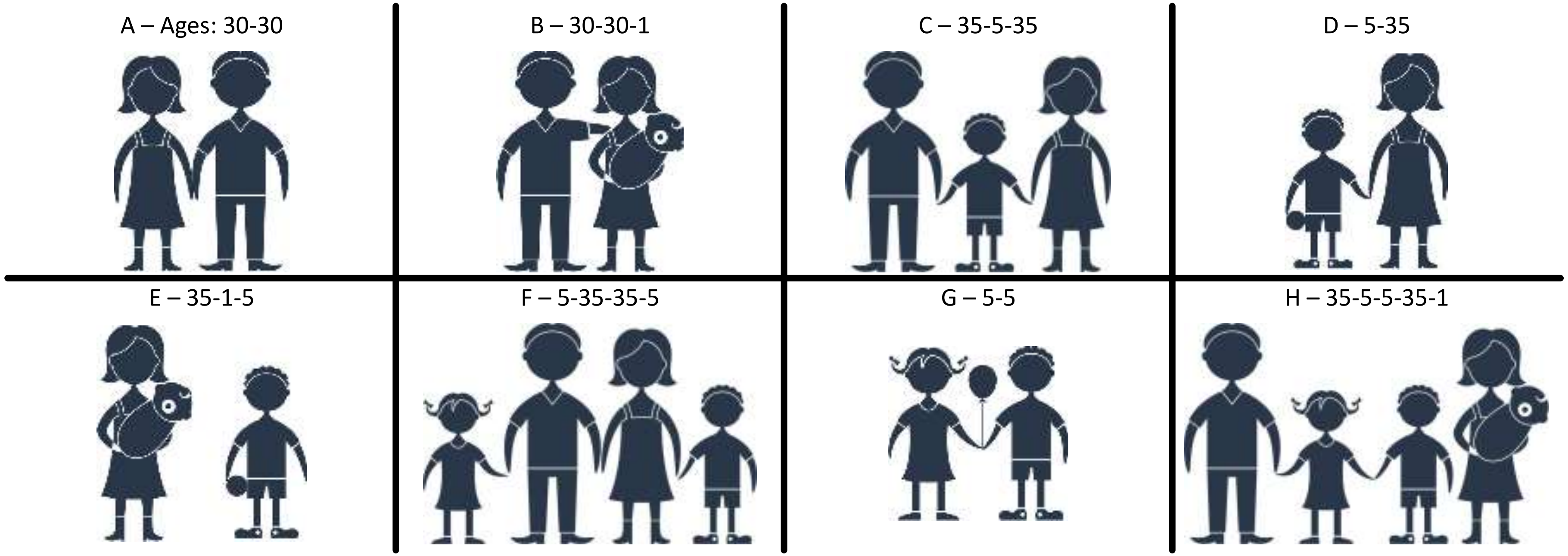
# Activity | How would you summarize this data? (cont.)

EXERCISE



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

EXERCISE



A – Ages: 30-30

B – 30-30-1

C – 35-5-35

D – 5-35

E – 35-1-5

F – 5-35-35-5

G – 5-5

H – 35-5-5-35-1

macrovector © 123RF.com

# Activity | Measures of Central Tendency. 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]<br>(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

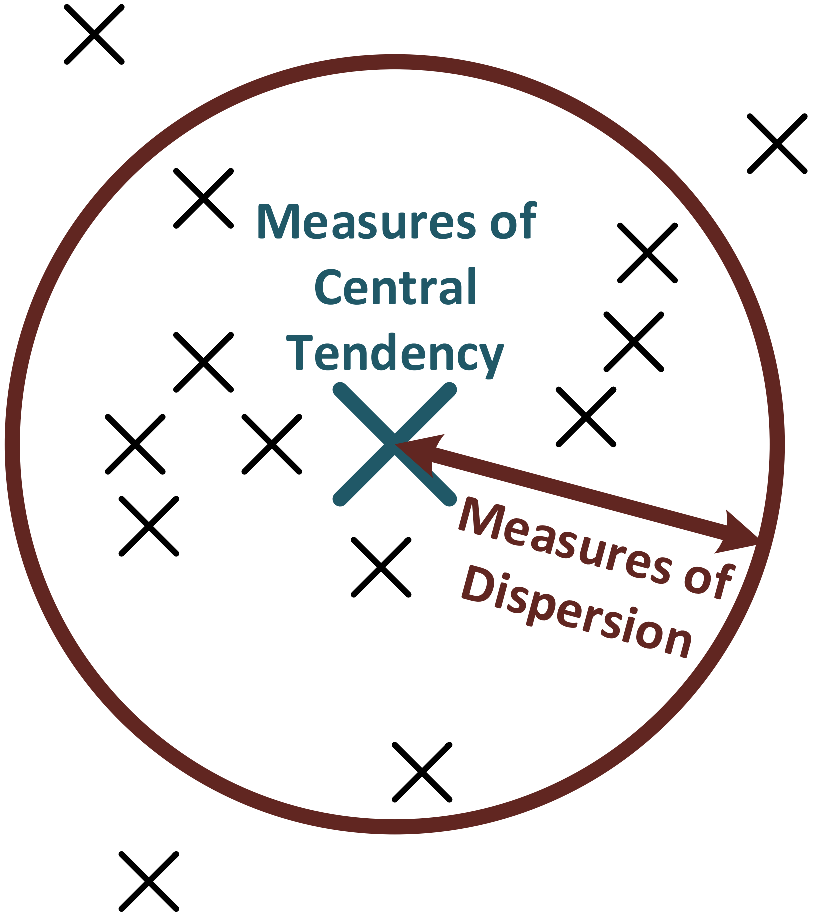
# Mean, Median, and Mode | Trade-offs

| | 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 | ☹<br>(Unlikely) | ☺ | ☹ | ☺<br>(Variance, Standard Deviation) | ☺ |
| Median | 😐<br>(50% chance) | 😐<br>(need to rank the values) | ☺ | ☺<br>(Interquartile Range) | ☹ |
| Mode | ☺<br>(Always) | ☹<br>(Need to count and rank the count) | ☺ | ☹<br>(Not really) | ☹<br>(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**<br>*(a.k.a., the first moment)*<br>(Mean has unit of $X$:$[X]$) | $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = E[X^1]$<br>(mu) | $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$<br>(x-bar) |
| **Variance**<br>*(a.k.a., the second moment)*<br>$[X^2]$ | $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$<br>$= E[(X - \mu)^2]$<br>(sigma-squared) | $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$ |
| **Standard Deviation**<br>$[X]$ | $\sigma = \sqrt{\sigma^2}$<br>(sigma) | $s = \sqrt{s^2}$ |

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

# ❸ PARSE the Data

*Codealong – Part B*

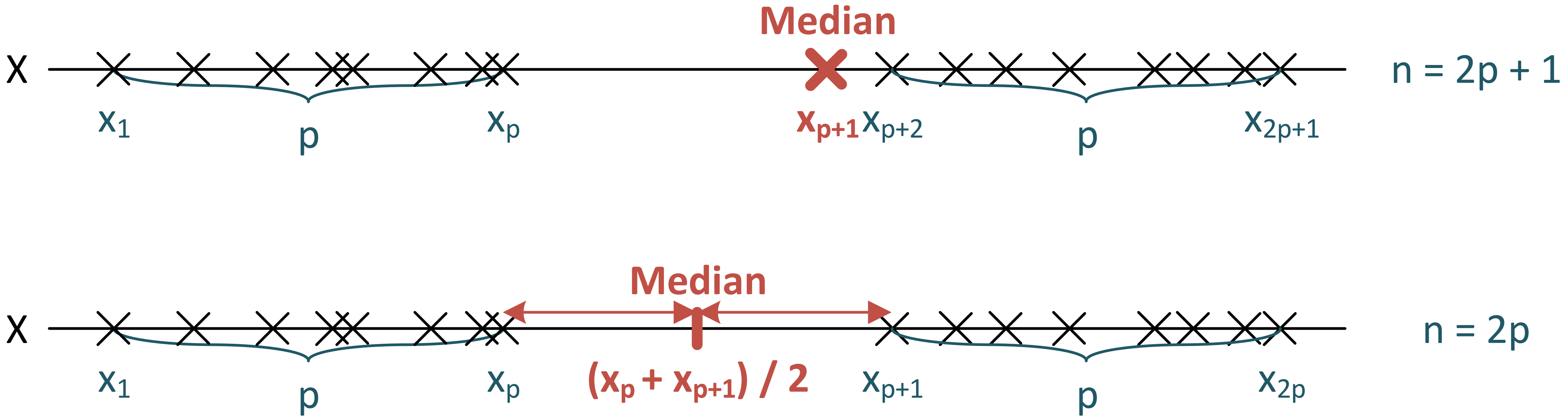
*`.mean()`*

*`.var(), .std()`*

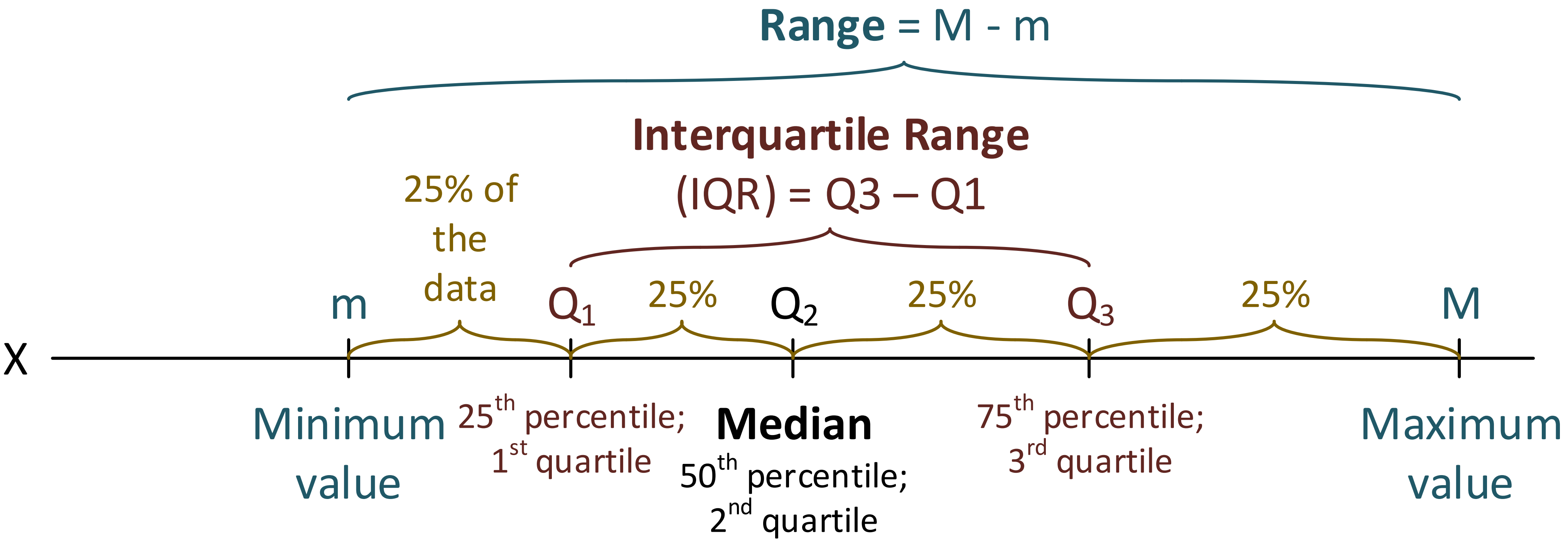
# ❸ PARSE the Data

*Median, Range, and Interquartile Range*

# Median


Median
X
$x_1$
p
$x_p$
$x_{p+1}$
$x_{p+2}$
p
$x_{2p+1}$
n = 2p + 1
Median
X
$x_1$
p
$x_p$
$(x_p + x_{p+1}) / 2$
$x_{p+1}$
p
$x_{2p}$
n = 2p

# Median, Range, and Interquartile Range



# Median, Range, and Interquartile Range (cont.)

| Nominal ✘ | Ordinal ✘ | Interval ✓ | Ratio ✓ |
|---|---|---|---|

| | |
|---|---|
| **Median** | $median = \begin{cases} x_{p+1} \; if \; n = 2p+1 \\ \frac{x_p + x_{p+1}}{2} \; if \; n = 2p \end{cases}$ |
| **Range** | $range = x_n - x_1$ |
| **Percentile** | $q_k = \begin{cases} x_{\lceil p \rceil} \; if \; p = \frac{nk}{100} \; not \; integer \\ \frac{x_p + x_{p+1}}{2} \; otherwise \end{cases}$ |
| **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

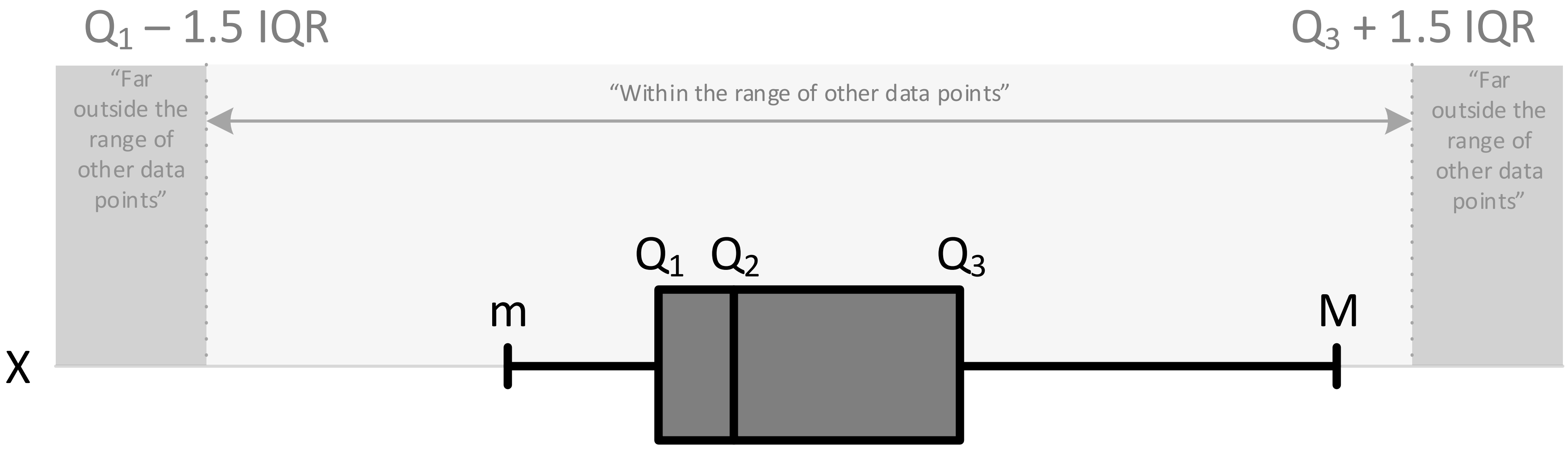
*Codealong – Part C*

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

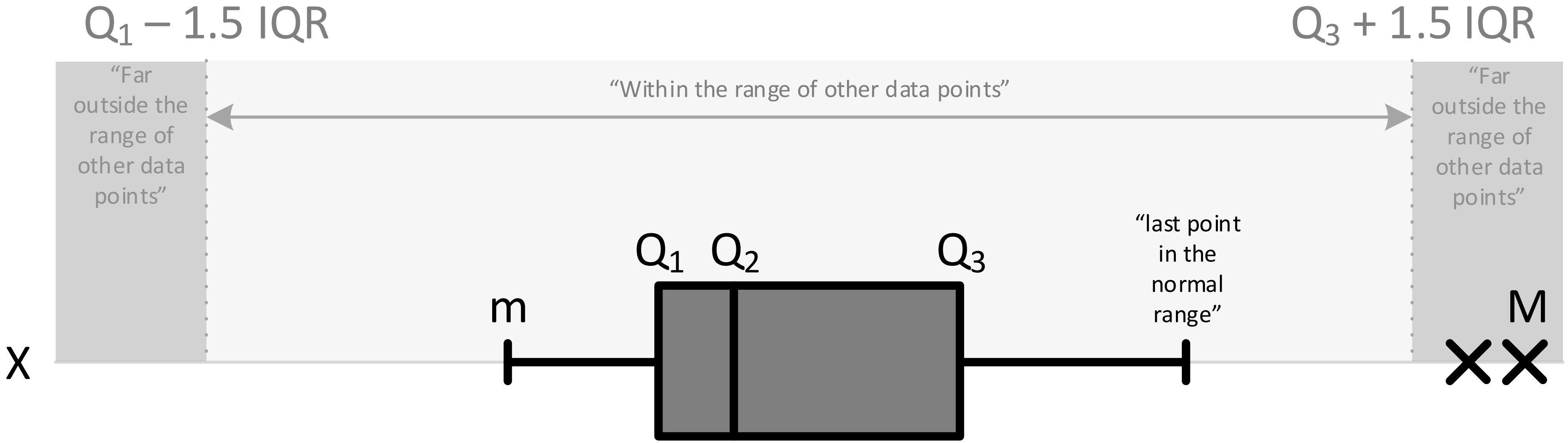
# ❸ PARSE the Data

*Median, Range, Interquartile Range, and Boxplots*

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


$Q_1$ – 1.5 IQR
$Q_3$ + 1.5 IQR
“Far outside the range of other data points”
“Within the range of other data points”
“Far outside the range of other data points”
$Q_1$
$Q_2$
$Q_3$
m
M
X

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


$Q_1$ – 1.5 IQR
$Q_3$ + 1.5 IQR
"Far outside the range of other data points"
"Within the range of other data points"
"Far outside the range of other data points"
"last point in the normal range"
$Q_1$
$Q_2$
$Q_3$
m
M
X

# ❸ PARSE the Data

*Codealong – Part D*

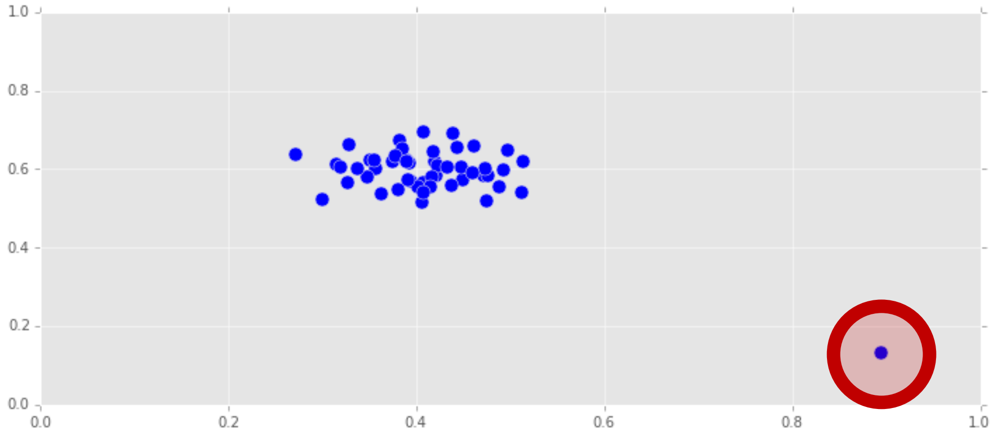
*Boxplots*

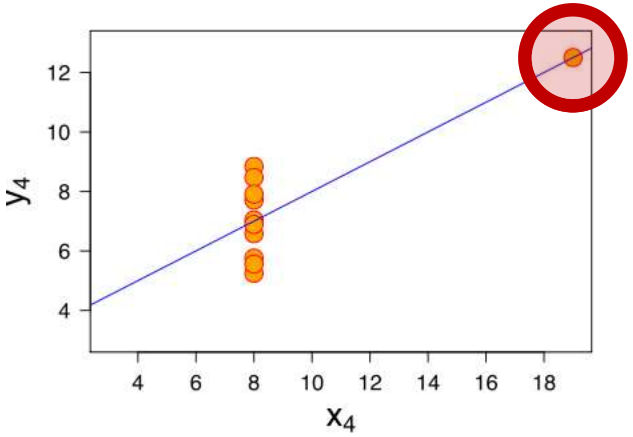
# ❸ PARSE the Data

*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

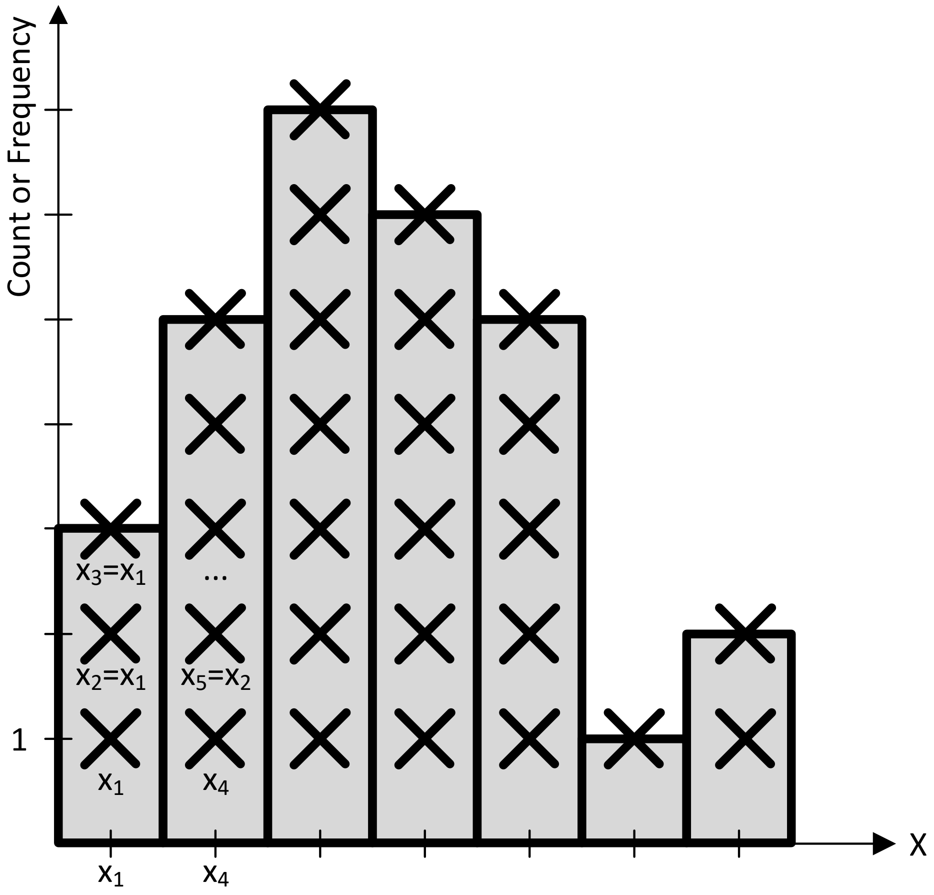




# ❸ PARSE the Data

*Histograms*

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



# ❸ PARSE the Data

*Codealong – Part E*

*Histograms*

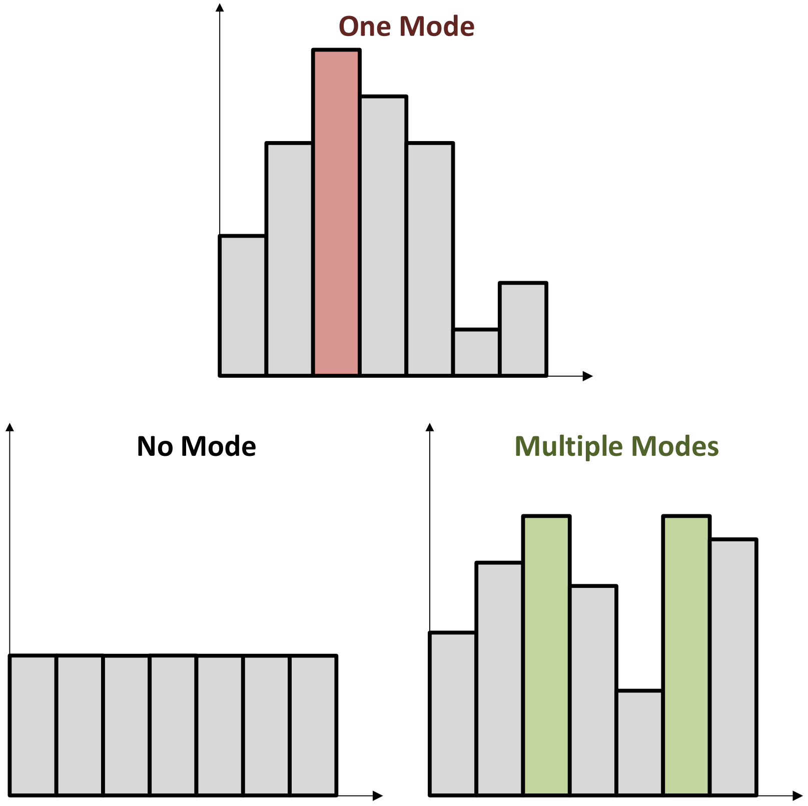
# ❸ PARSE the Data

*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 F*

*`.mode()`*

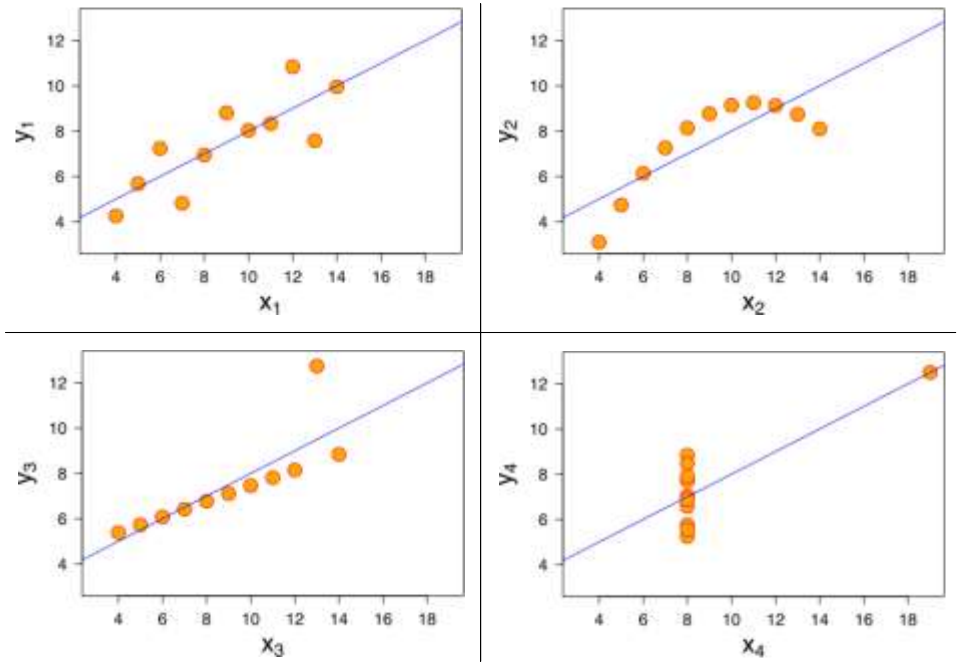
# ❸ PARSE the Data

*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.

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.



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.

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

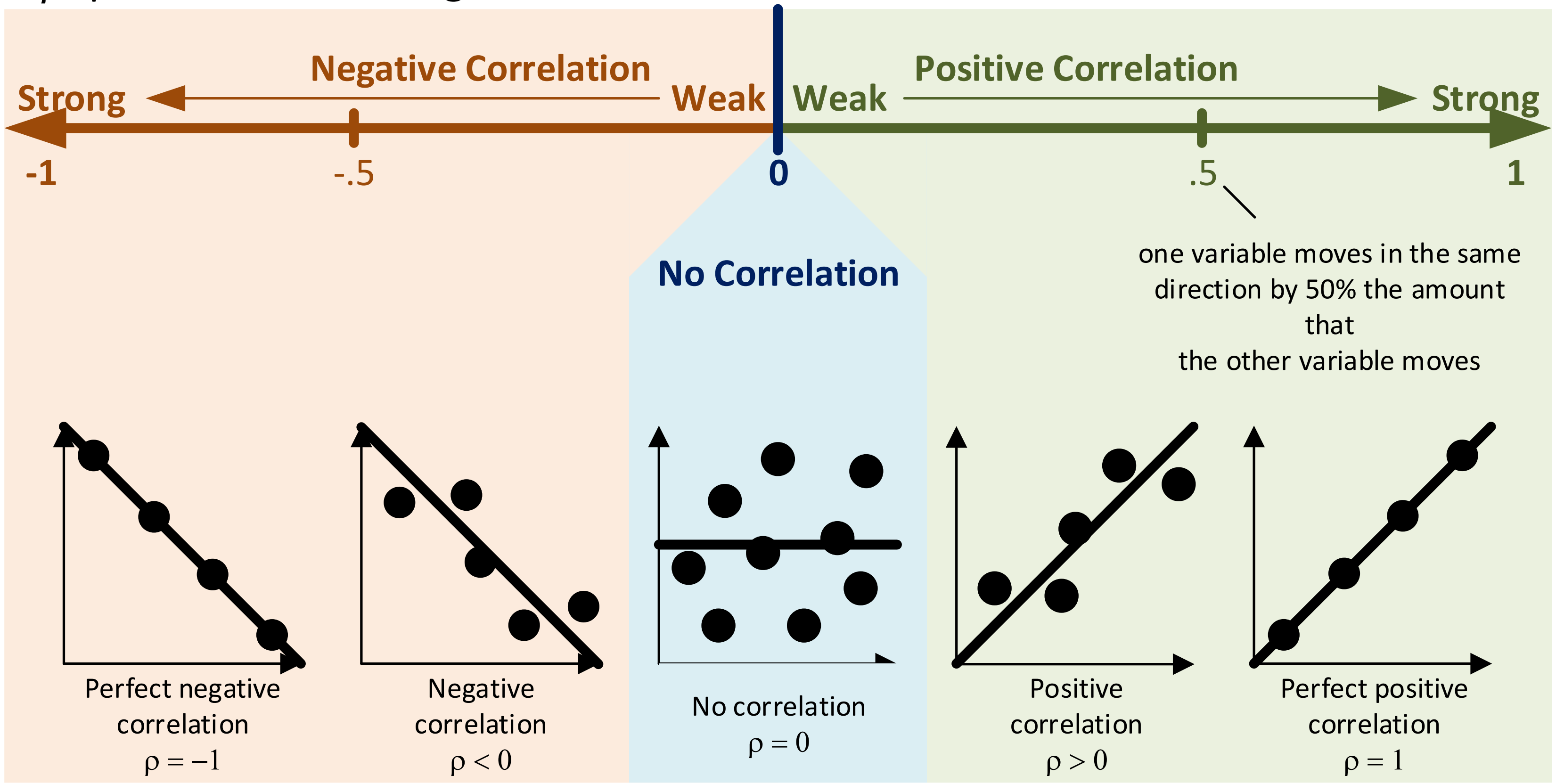
- A measure of strength and direction for a **linear association** between two random variables

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

- ρ = 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



one variable moves in the same direction by 50% the amount that
the other variable moves

Perfect negative correlation $\rho = -1$

Negative correlation $\rho < 0$

No correlation $\rho = 0$

Positive correlation $\rho > 0$

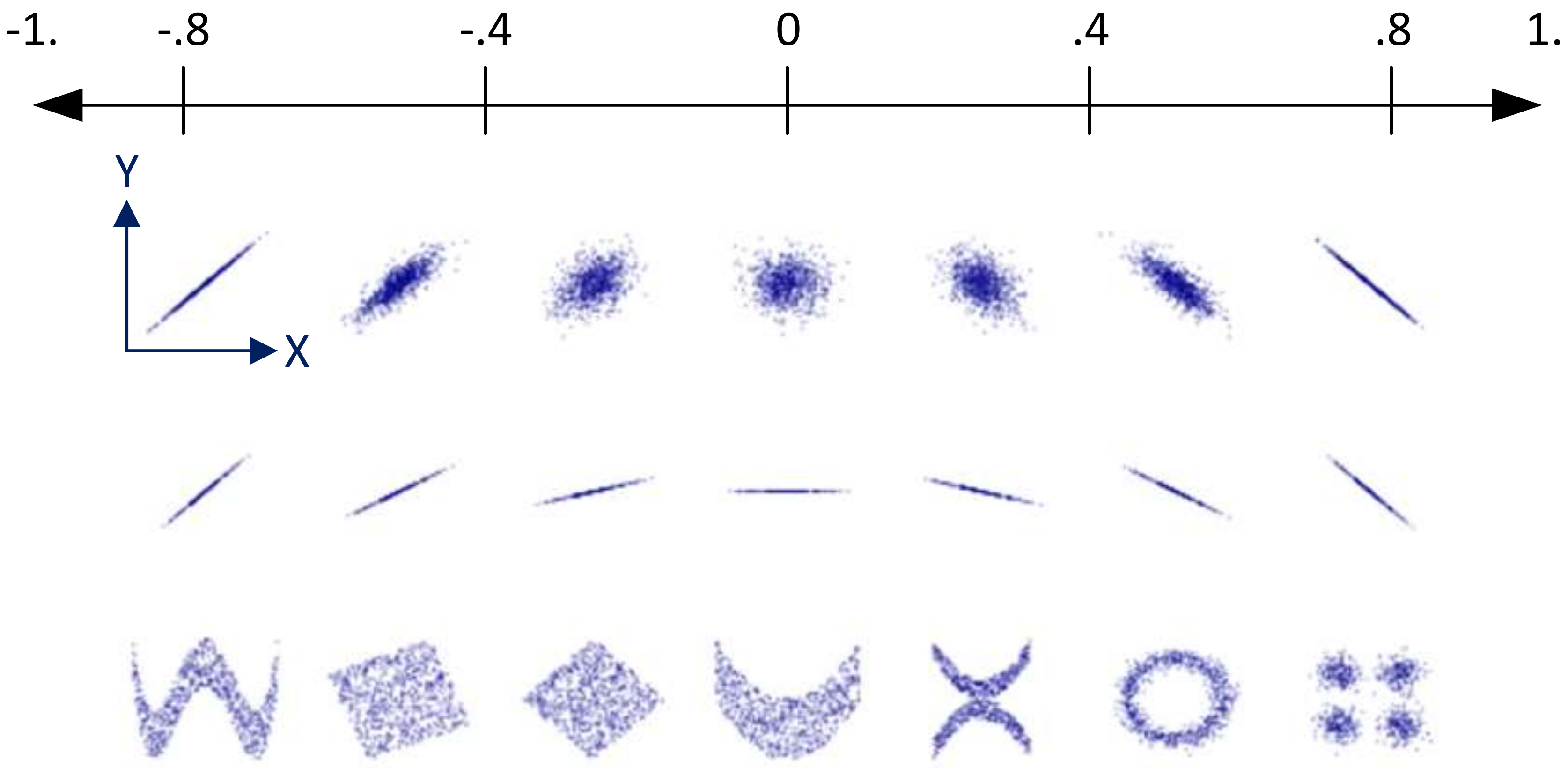
Perfect positive correlation $\rho = 1$

# ❸ PARSE the Data

*Activity | Correlations and Scatter Plots*

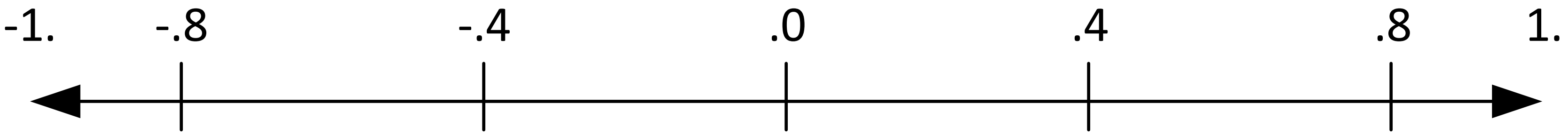
# Activity | What’s the correlations for the following scatter plots (5 minutes)

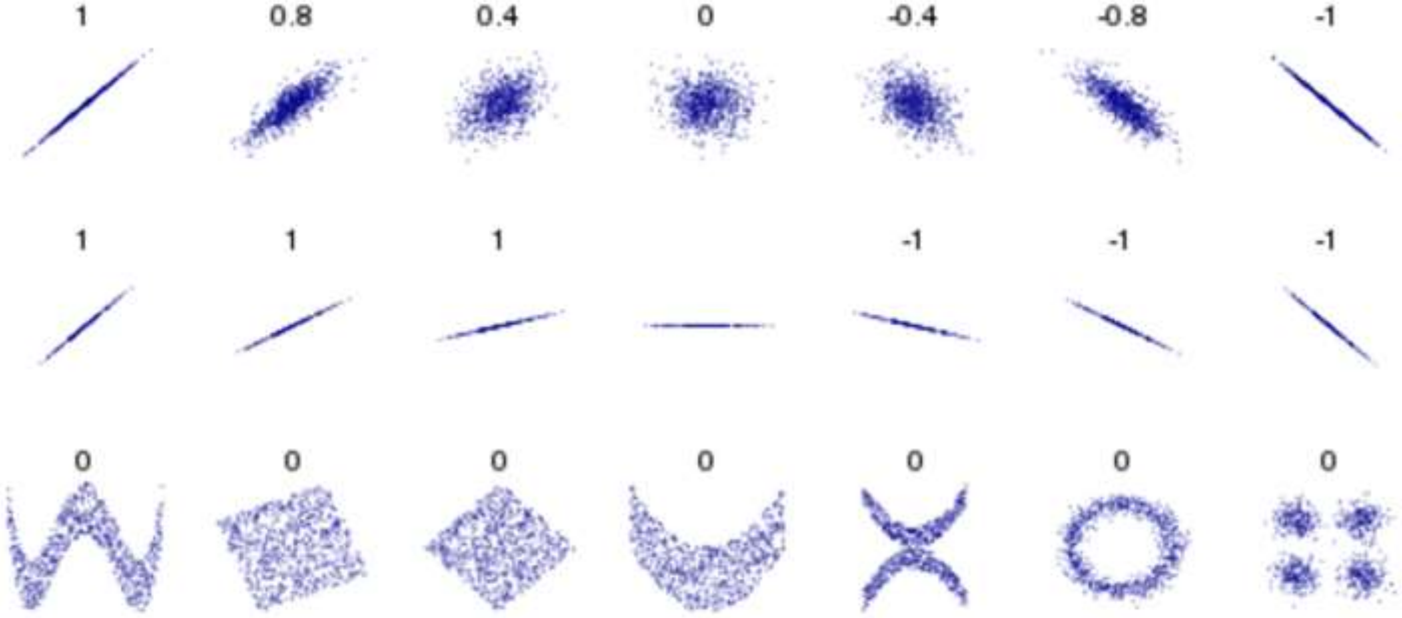
EXERCISE



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

EXERCISE





Source: Wikipedia

# ❸ PARSE the Data

*Codealong – Part G*

*`.corr()`*

*Heatmaps*
*Scatter plots and matrices*

# ❸ PARSE the Data

*Codealong – Part H*

*`.value_counts()`*

*`.crosstab()`*

# Lab

*Exploratory Data Analysis*

# 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

# Next Class

*Flexible Class Session #1 | Exploratory Data Analysis*

# Exit Ticket

*Don’t forget to fill out your exit ticket here*