# Statistics Fundamentals

*Ivan Corneillet*

*Data Scientist*

# Learning Objectives

After this lesson, you should be able to:

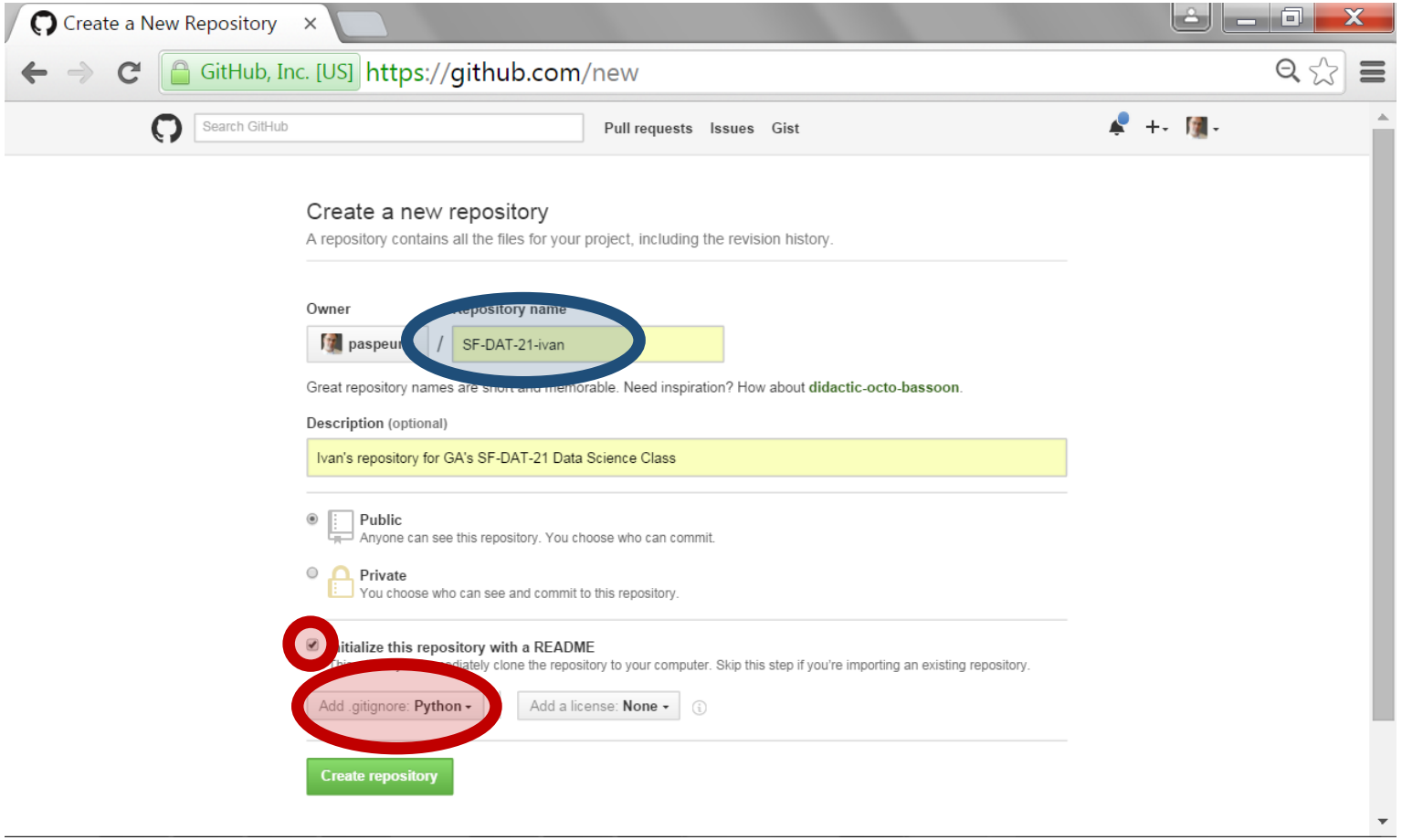
- ID variable types
- Use the *pandas* (and *NumPy*) libraries to analyze datasets using basic summary statistics: mean, median, mode, max, min, quartile, inter-quartile range, variance, standard deviation, and correlation
- Create data visualizations – including: boxplots, histograms, and scatter plots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

# Outline

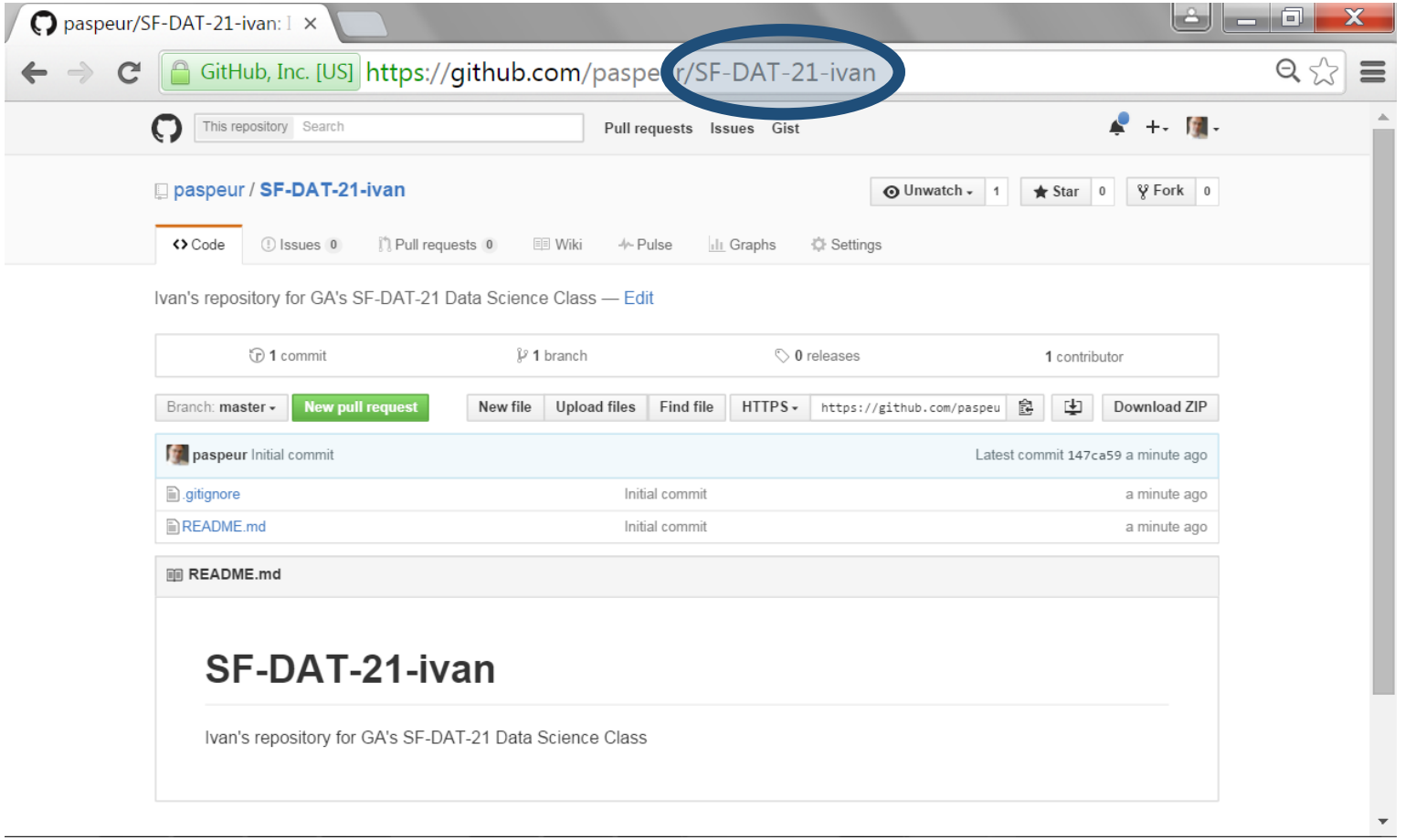
- Unit Project 1 due today
- Submitting Your Work (via the GitHub web interface)
- Review
- ❸ Parse the Data
  - Types of Data and Types of Measurement Scales
  - Populations and Samples; Descriptive vs. Inferential Statistics
  - Measures of Central Tendency and Measures of Dispersion
  - Boxplots
  - Outliers
  - Histograms
  - Measurement Errors
  - Outliers
  - Histograms
  - Measurement Errors
  - Correlation
- Review
- Assigned
  - Unit Project 2 (due in 1 week)
- In-flight
  - Final Project 1 (due in 2.5 weeks)


# Submitting Your Work (via the GitHub web interface)

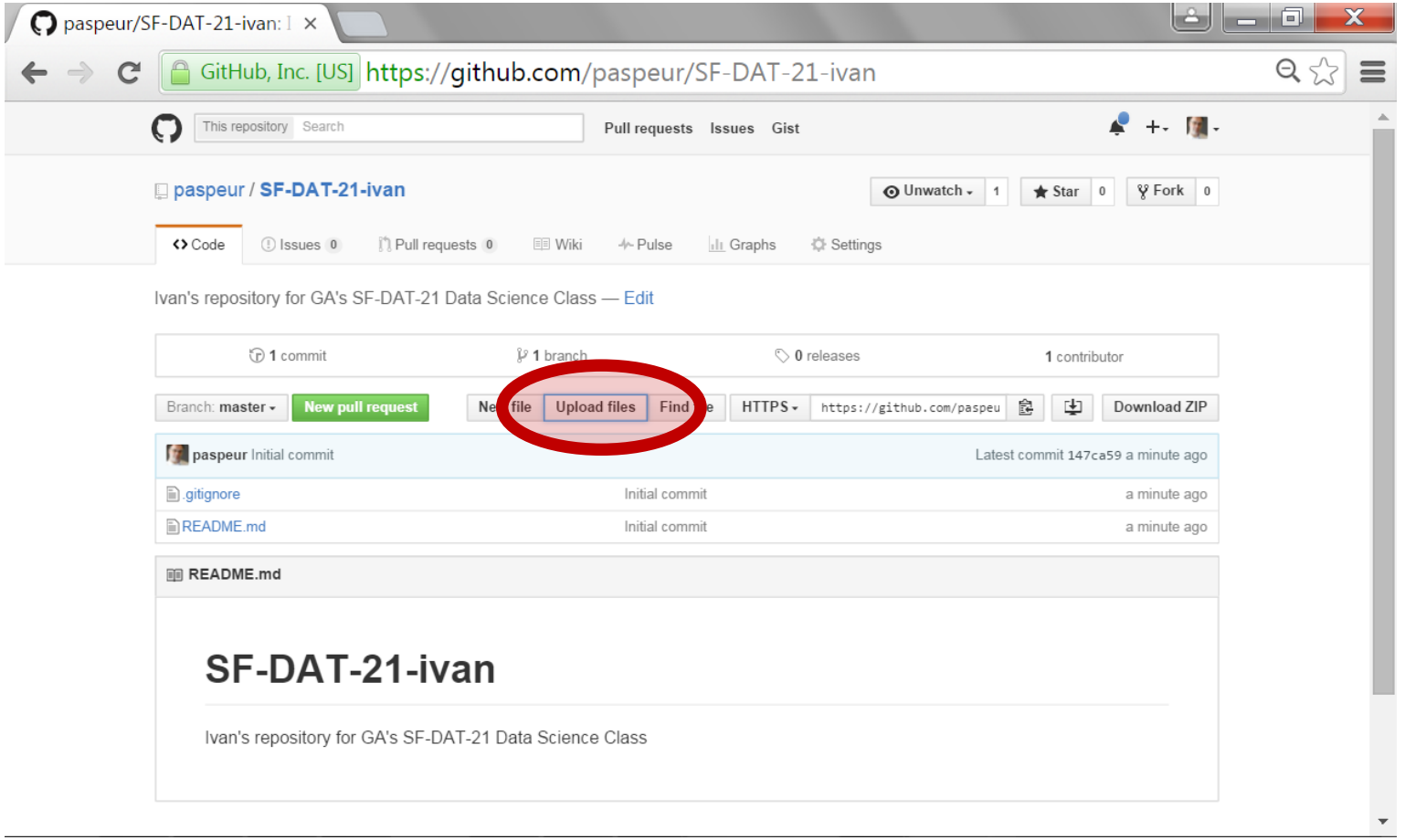
1. You need to create your own GitHub repo to submit your work (https://github.com/new). Select “Initialize this repository with a README” and select “Python” for “Add .gitignore”



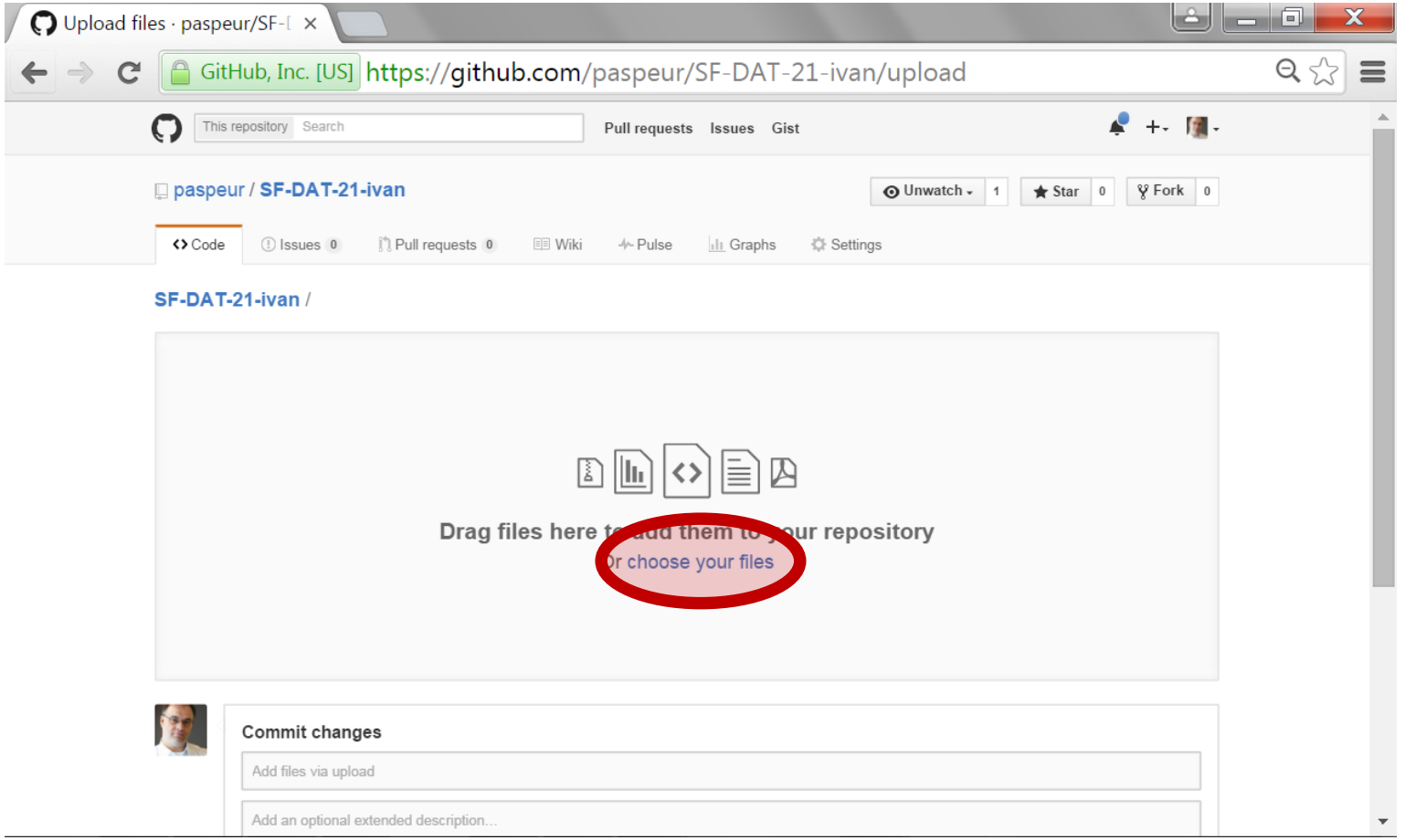
# 1. Done: Your new GitHub repo (cont.)



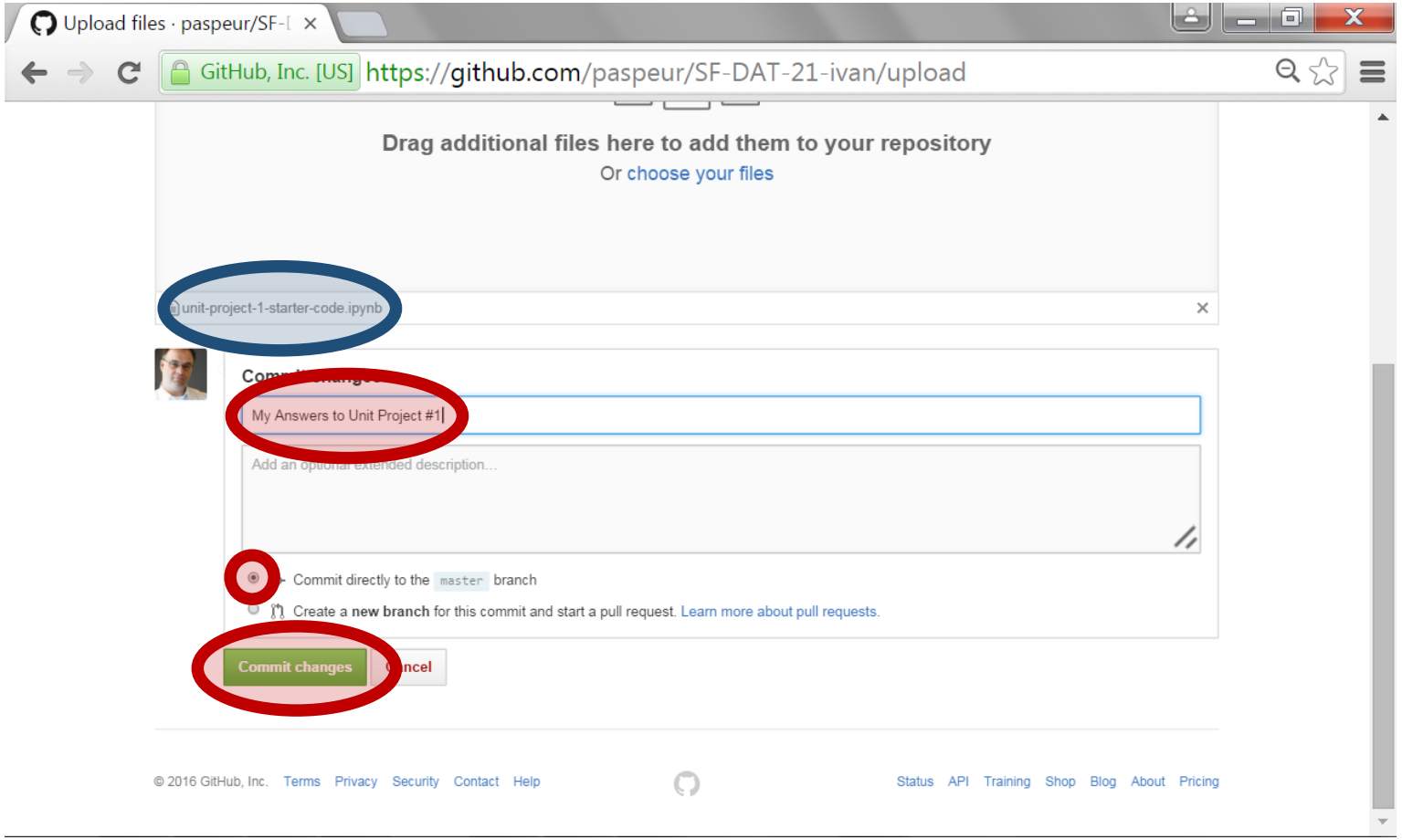
# 2a. Then click on `Upload File`



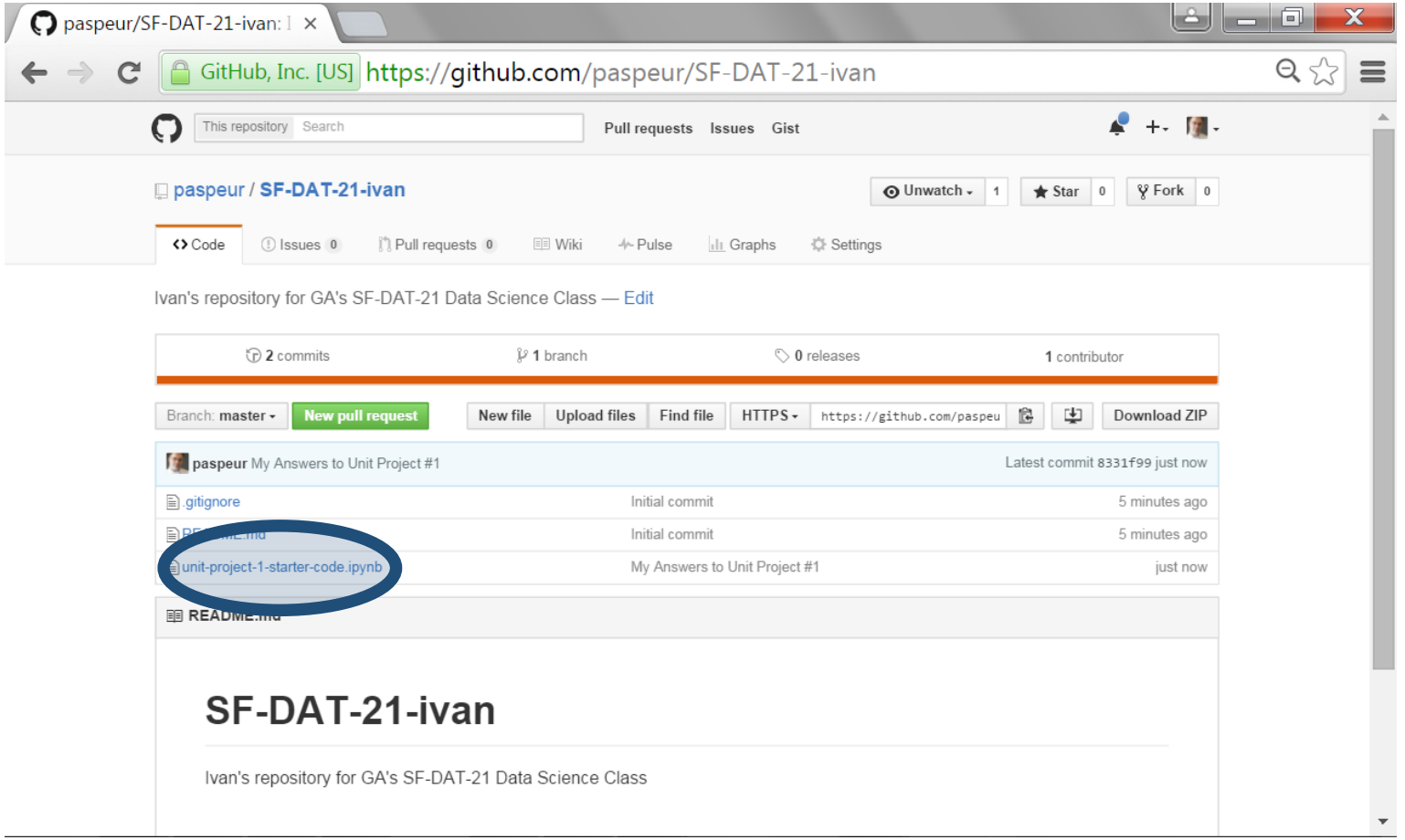
# 2b. Drag or choose your files



# 2c. Add a commit message and leave “Commit to the master branch” unchanged

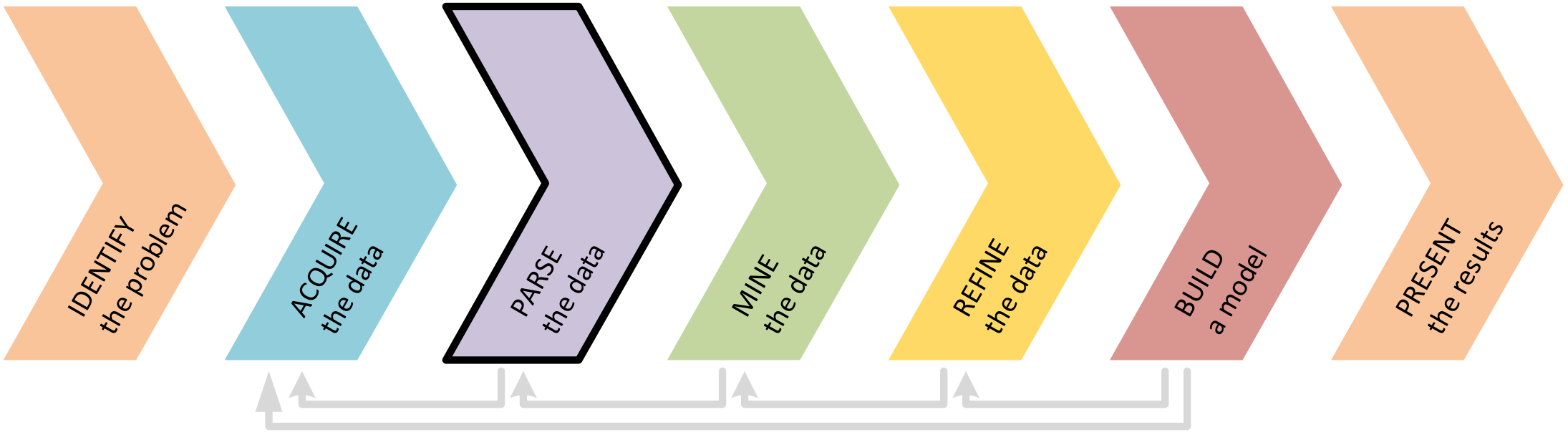


# 2d. Done! But learning to do it on the command line is cool too!




# Review and Activity
# Data Science Workflow

# Today we'll keep our focus on PARSE the data




# Review

❶ IDENTIFY the problem

❷ ACQUIRE the data

# Review

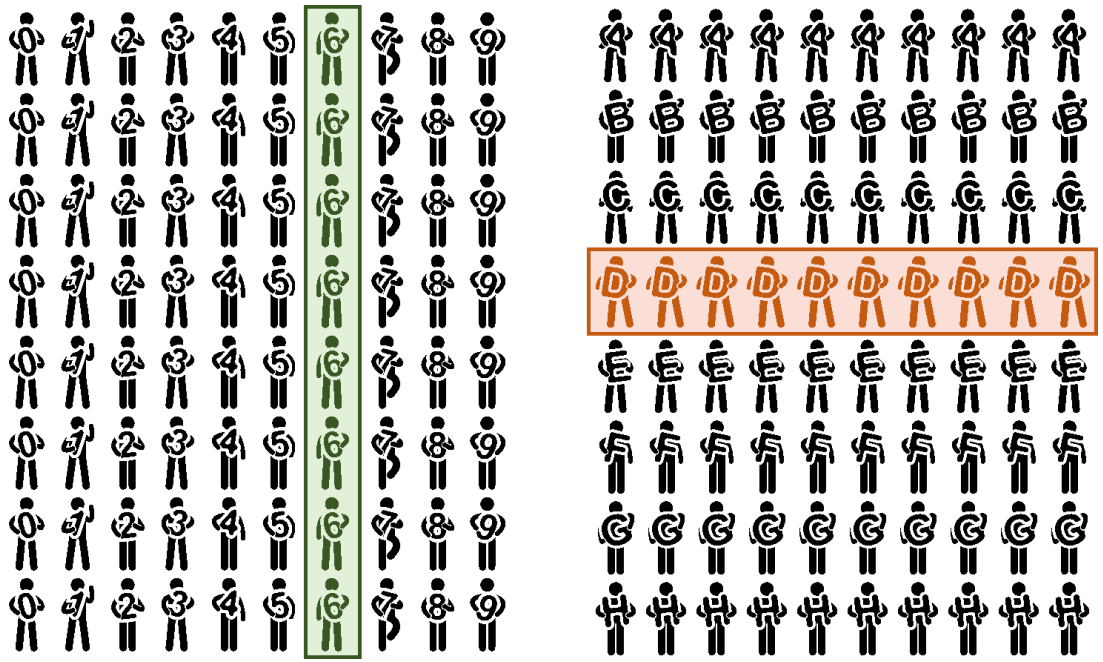
## ❶ IDENTIFY the problem

*SMART Goals*

| | |
|---|---|
| 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 |

## ❷ ACQUIRE the data

*Cross-sectional vs. Longitudinal Data*

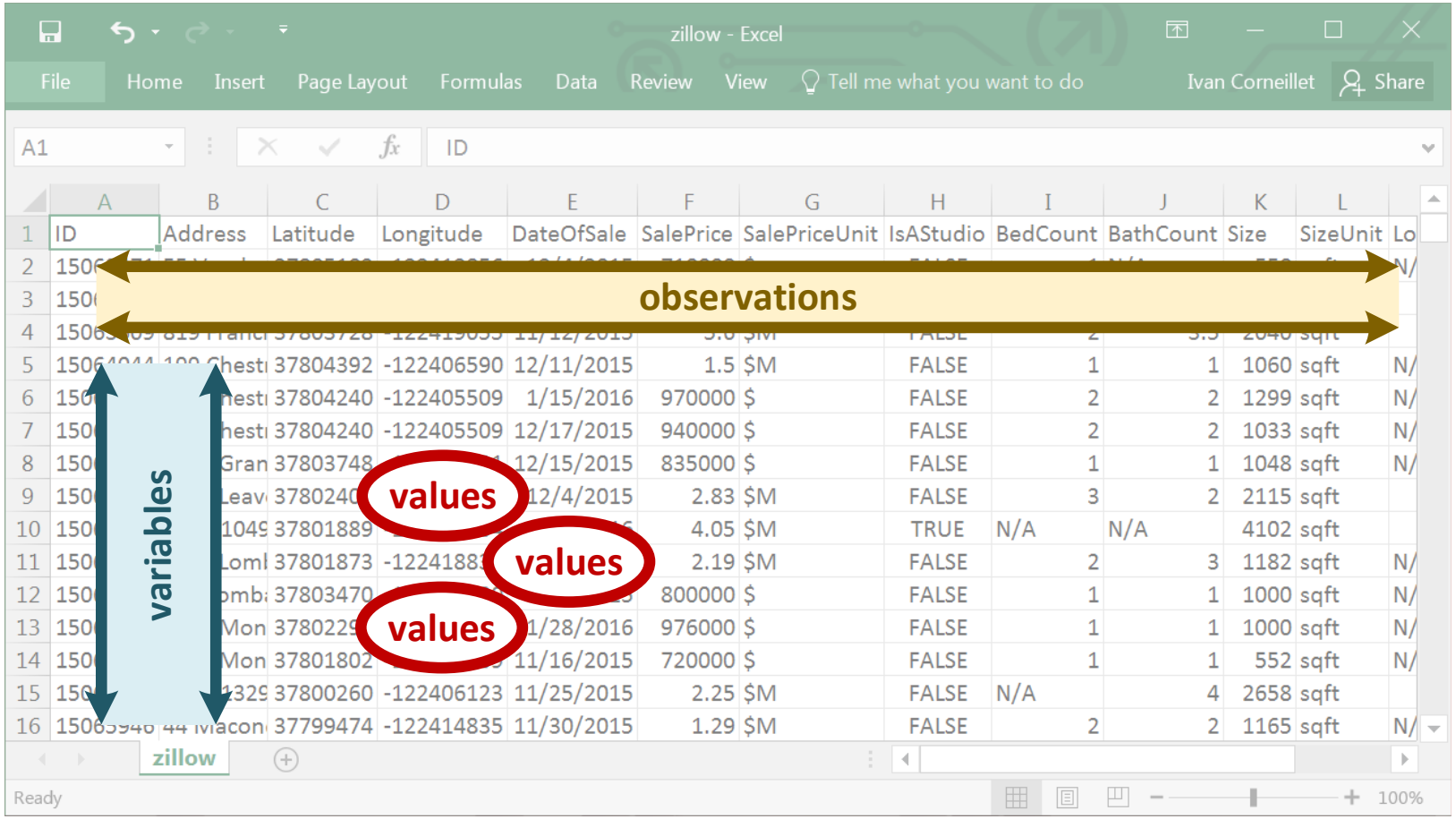


Khoon Lay Gan © 123RF.com


# Review
# ❸ Parse the Data

# ❸ Parse the Data

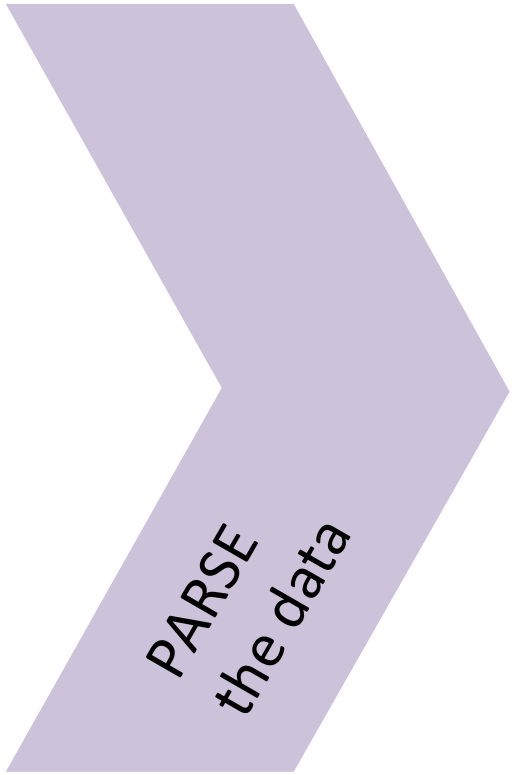
## *Tidy data and pandas*




# Q & A


# ❸ Parse the Data (cont.)

# 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)

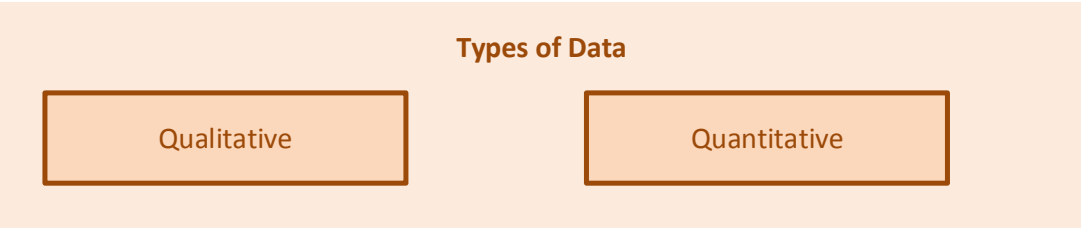
# ❸ Parse the Data (cont.)

- Parse the Data
  - Read any documentation provided with the data
  - Perform exploratory data analysis
  - Verify the quality of 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
- Measurement Errors
- Correlation

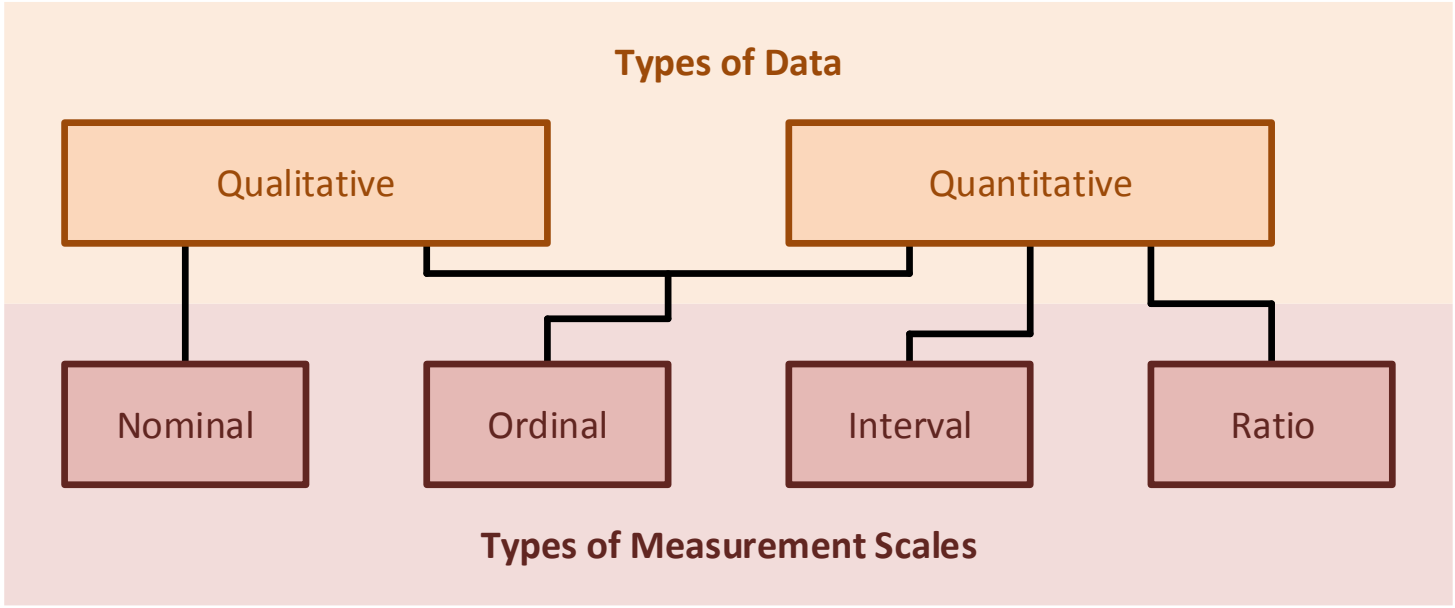

# Types of Data and Types of Measurement Scales

# 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 ☹) |


# Activity: Knowledge Check

# Activity: Knowledge Check

EXERCISE

## ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. What type of data are the columns in the Zillow dataset?
    a. Zillow ID
    b. Address
    c. Date of Sale
    d. Sale Price
    e. Whether it is a Studio
    f. Number of beds
    g. Number of baths
    h. Size
    i. Lot Size
    j. Year it was built

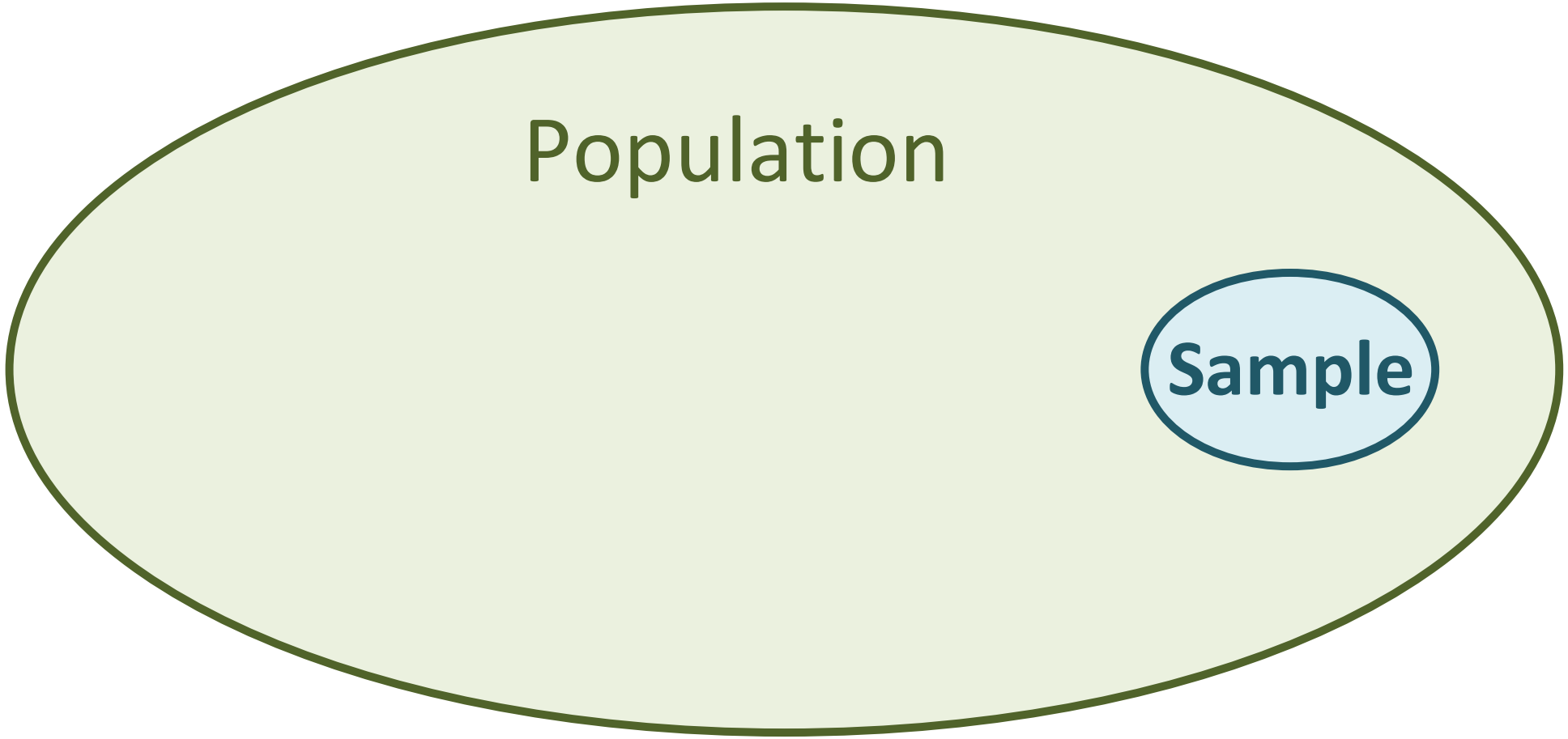
2. When finished, split into pairs and share your answers with each other

## DELIVERABLE

Answers to the above questions

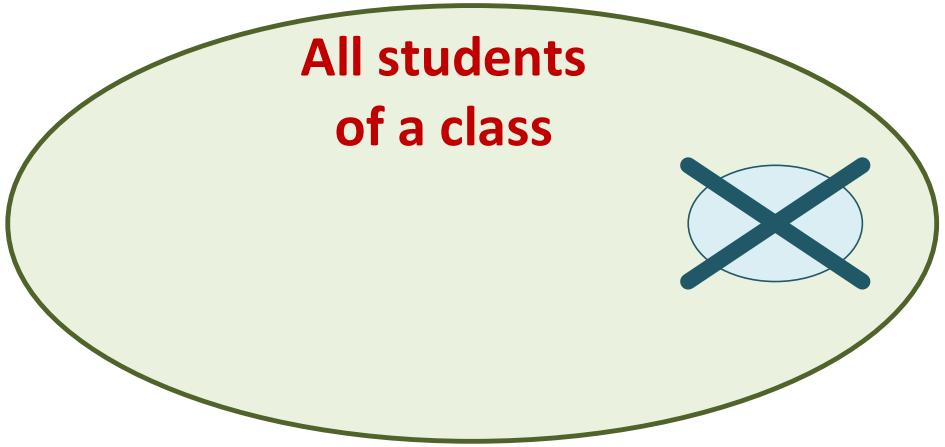

# Populations and Samples

# Populations and Samples

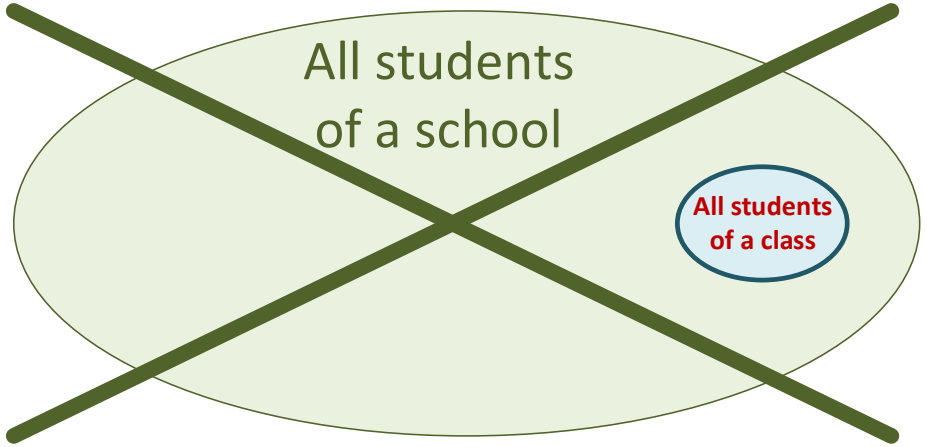


# A dataset may be considered either as a population or a sample, depending on the reason for its collection and analysis

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




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

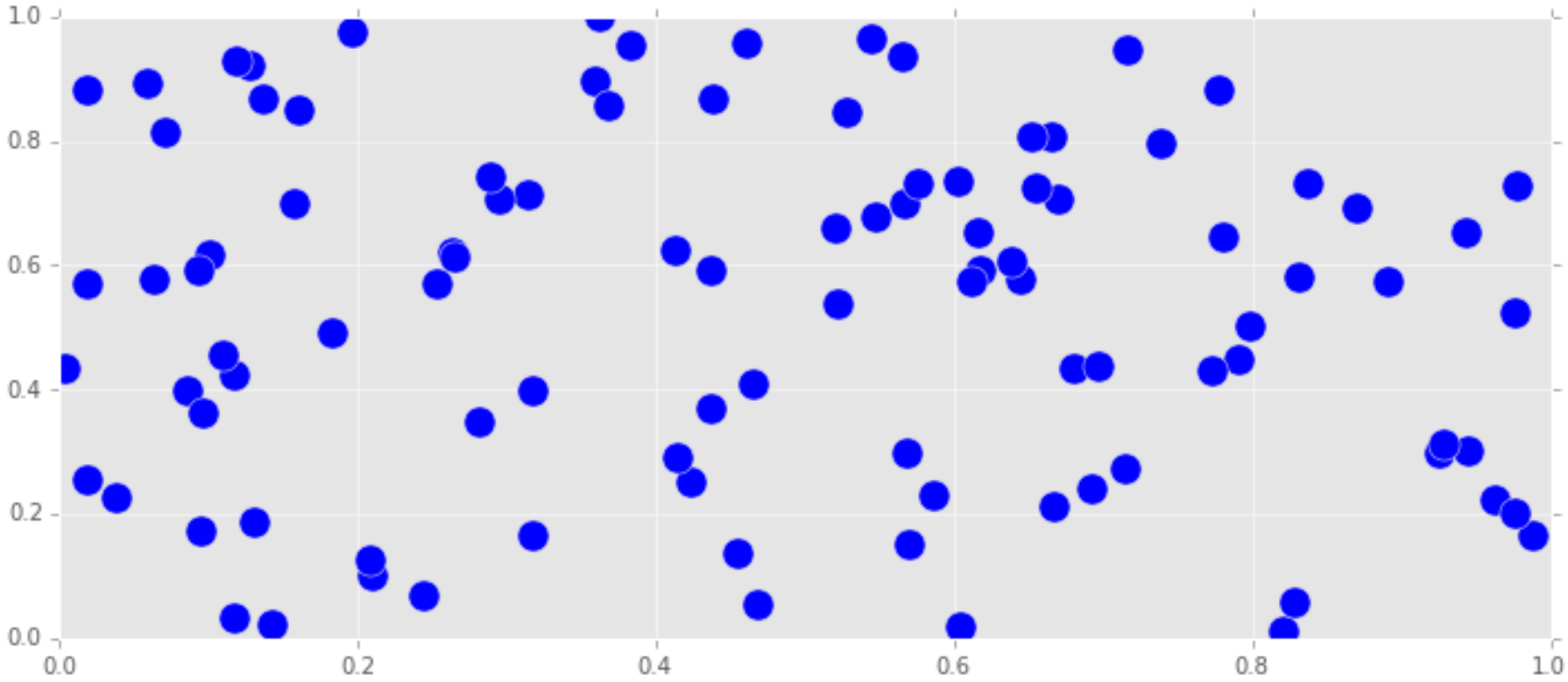



# Measures of Central Tendency and Measures of Dispersion


# Activity: Summaries and Measures of Central Tendency

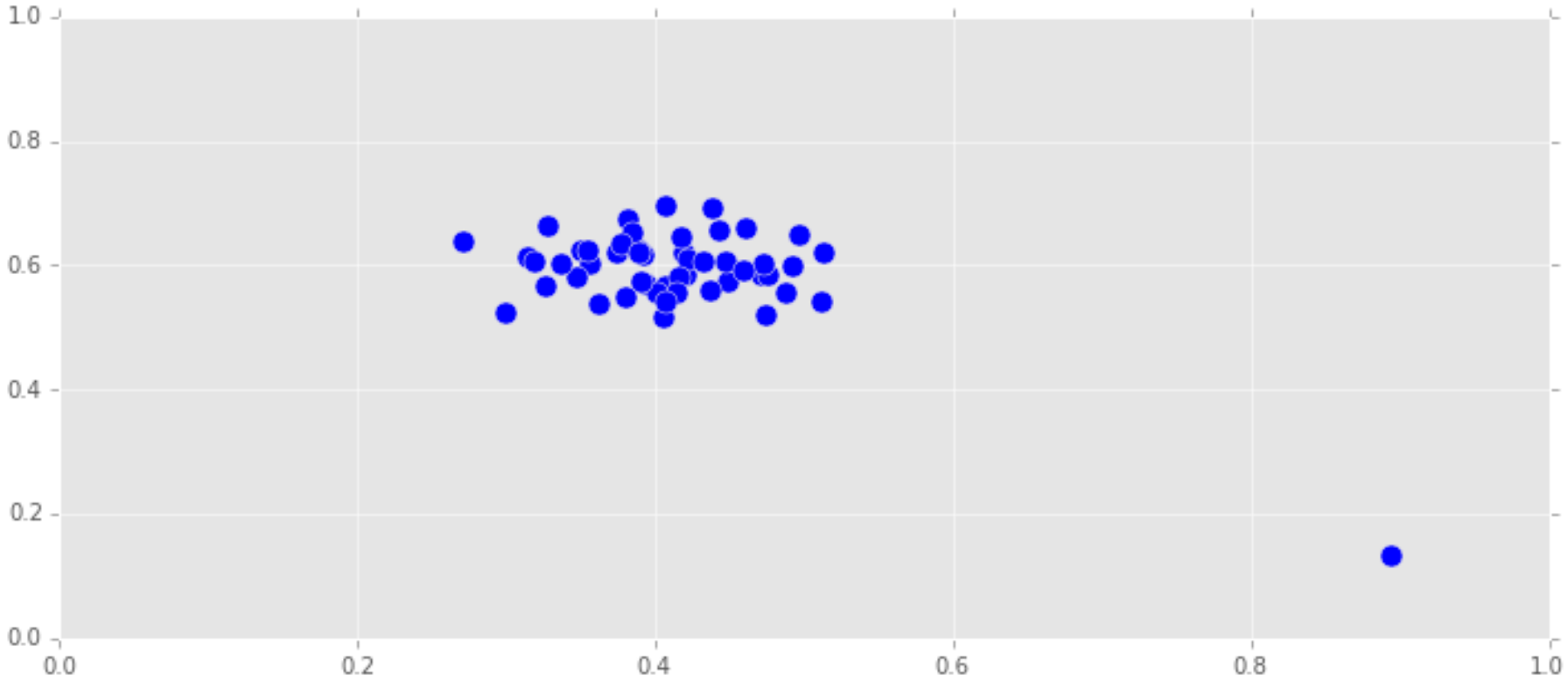
# Activity: How would you summarize this data?

**EXERCISE**



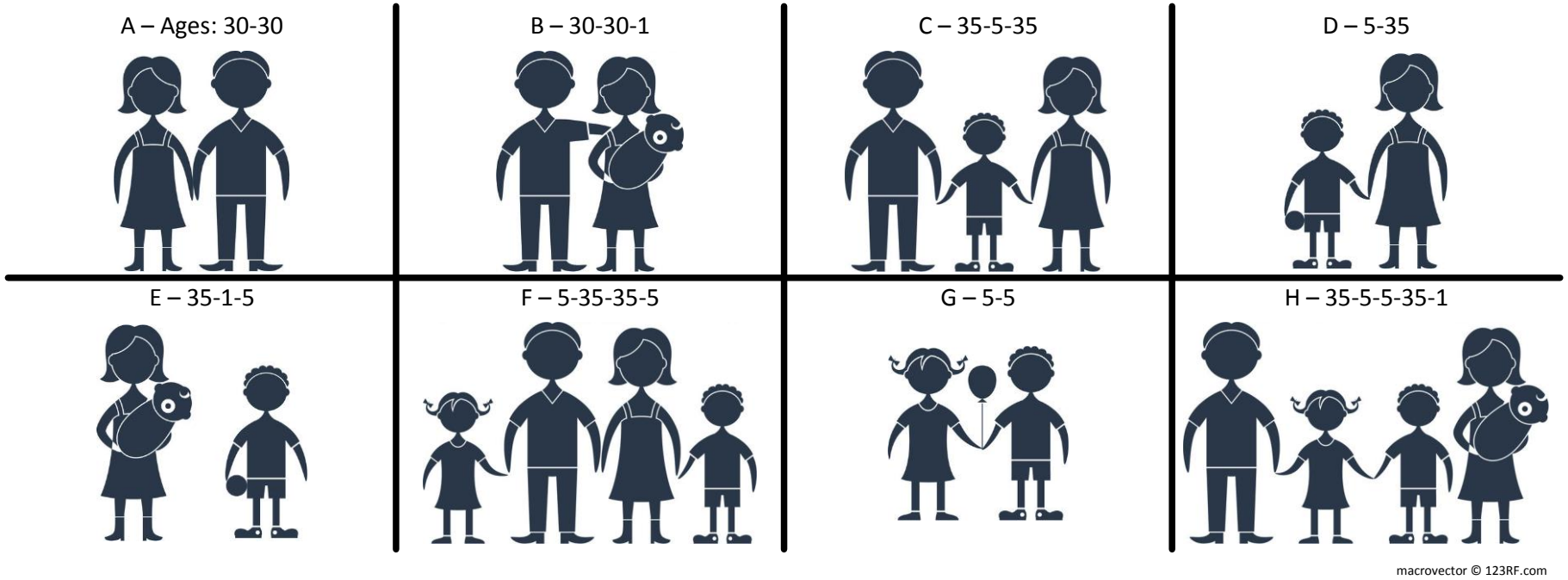
# Activity: How would you summarize this data? (cont.)

EXERCISE



# Activity: Measures of Central Tendency. What is the typical age for these 8 groups of people? (5 minutes)

**EXERCISE**



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

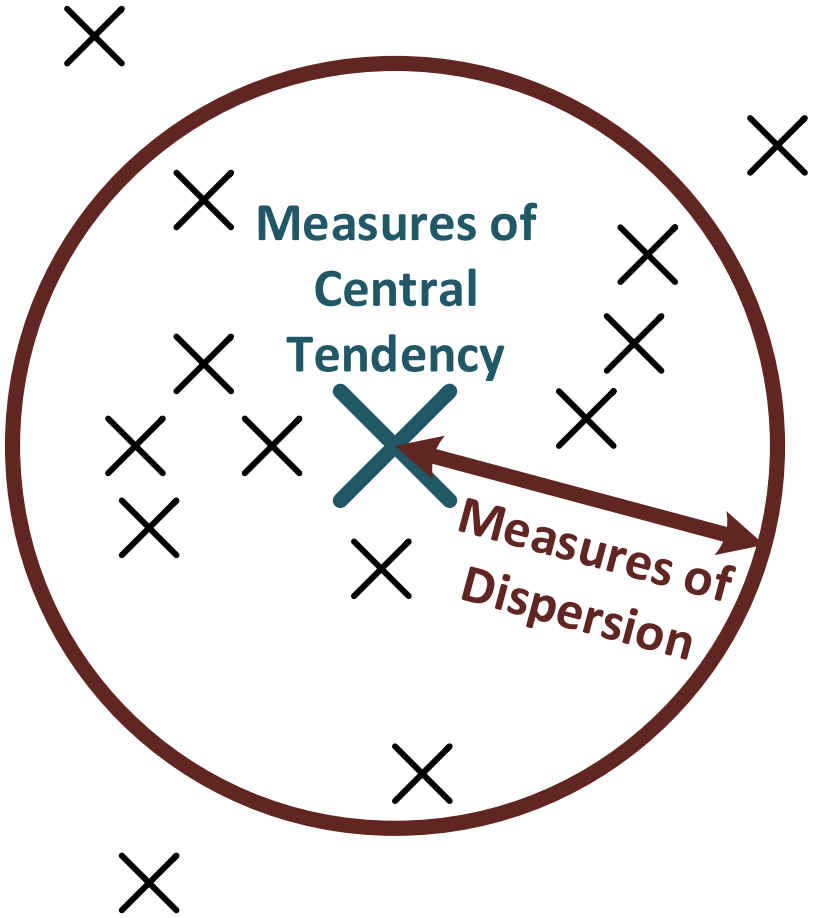
| Group | Mean | Median | Mode |
|---|---|---|---|
| **A** (30-30) | 30[1] | 30[1] | 30[1] |
| **B** (30-30-1) | 20.3[2] (i.e., no 20-year-olds in the group) | 30[3] | 30[3] |
| **C** (35-5-35) | 25[2] | 35[3] | 35[3] |
| **D** (5-35) | 20[2] | 20[2] | None[4] |
| **E** (35-1-5) | 13.6[2] | 5[2] | None[4]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

# There are no “Winners-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) |

# 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 $x_i$)

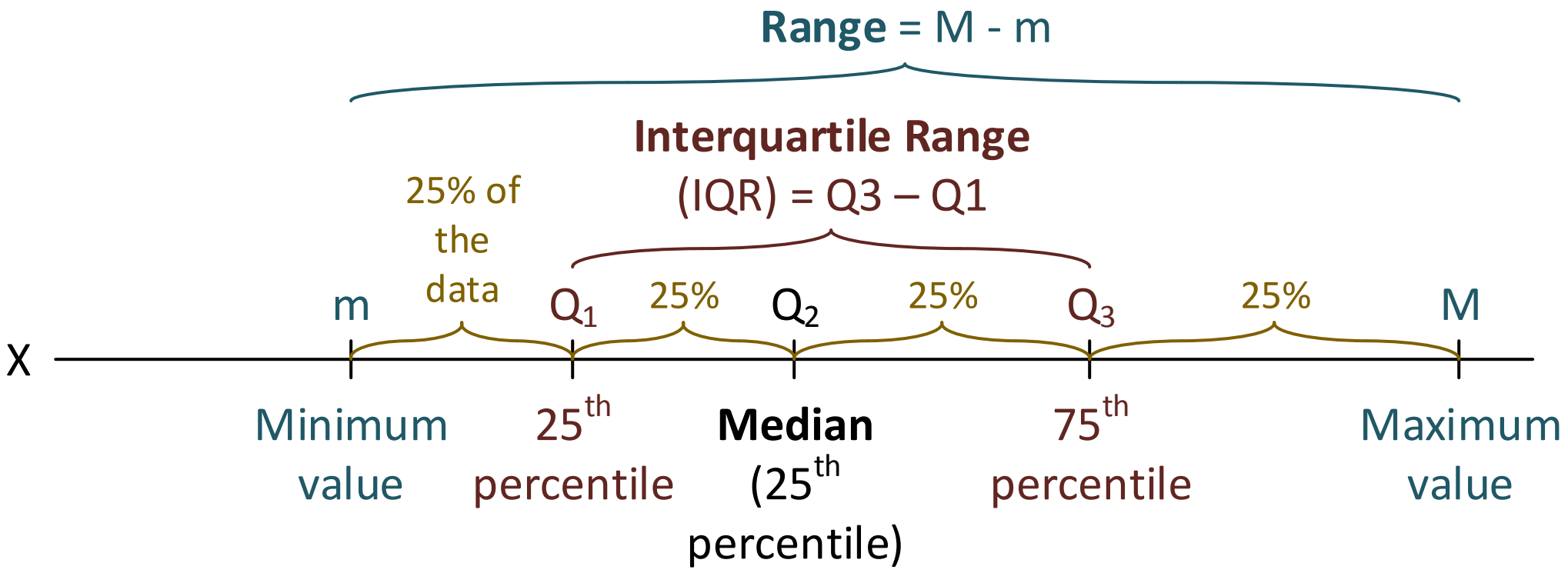

# Codealong: Part A

`.mean()`

`.var(), .std()`


# Median, Range, and Interquartile Range

# Median, Range, and Interquartile Range



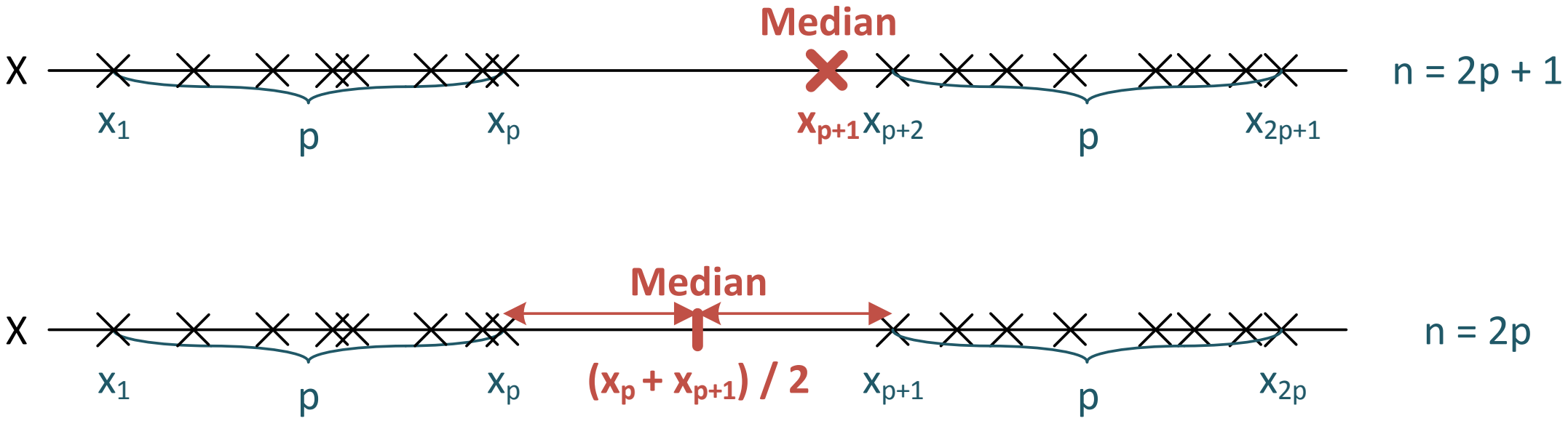
# Median, Range, and Interquartile Range (cont.)

| 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** | $IQ = Q_3 - Q_1$ |

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

# 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


# Codealong: Part B

`.median()`

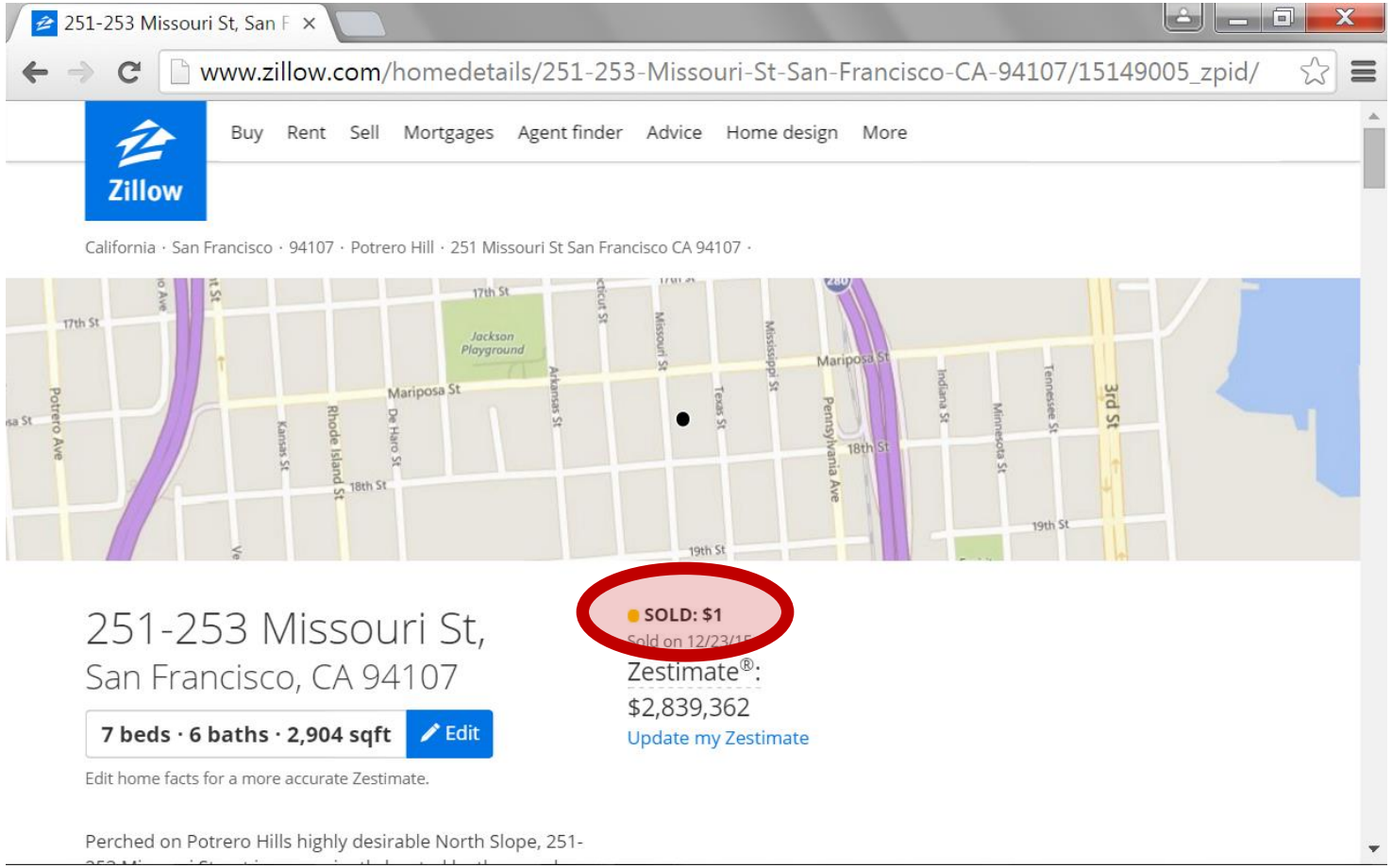
`.count()`, `.dropna()`, `.isnull()`

`.min()`, `.max()`

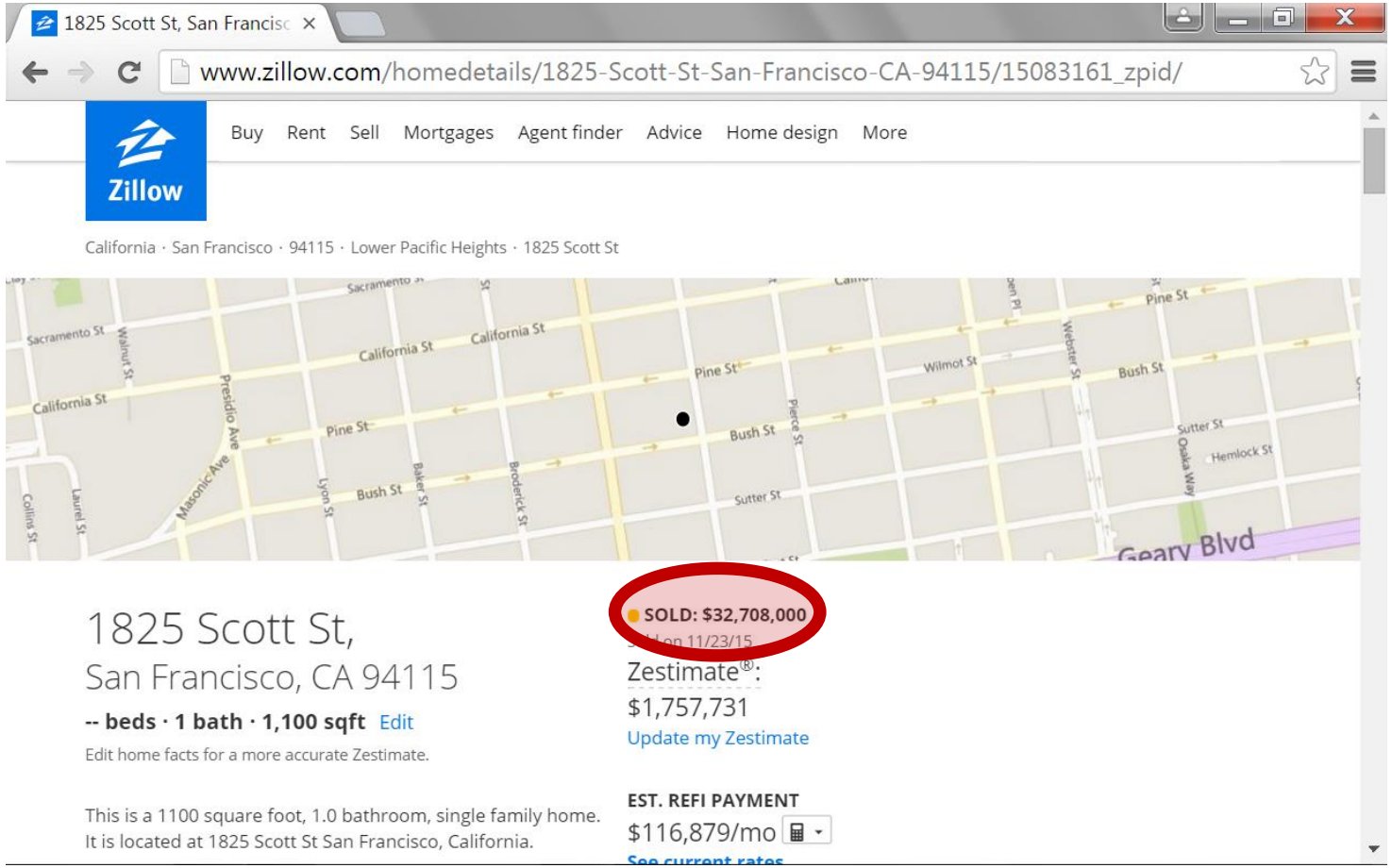
`.quantile()`

`.describe()`

http://www.zillow.com/homedetails/251-253-Missouri-St-San-Francisco-CA-94107/15149005_zpid/

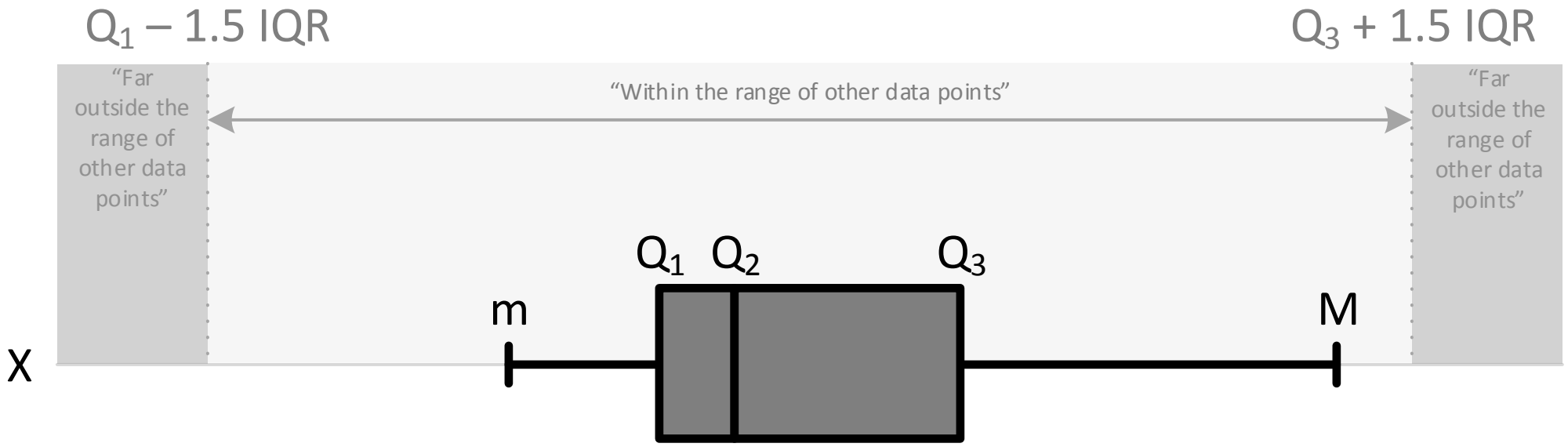


http://www.zillow.com/homedetails/1825-Scott-St-San-Francisco-CA-94115/15083161_zpid/

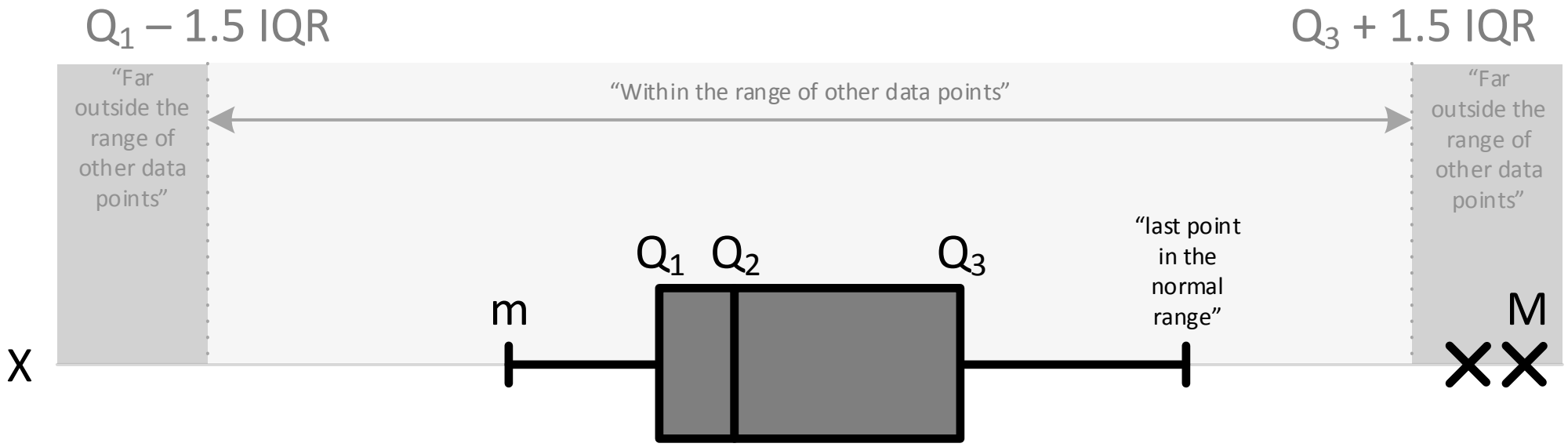



# Median, Range, Interquartile Range, and Boxplots

# Median, Range, Interquartile Range, and Boxplots



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

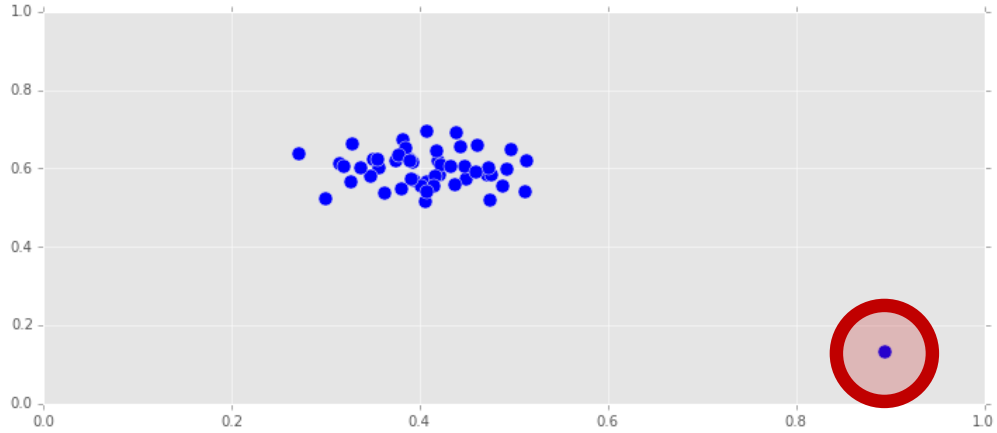


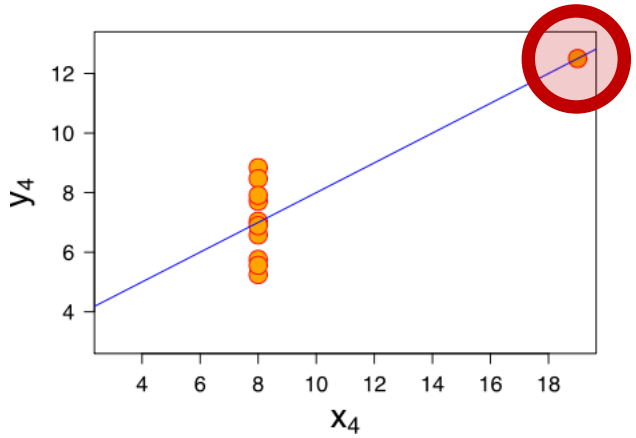

# Codealong: Part C
# Boxplots


# Outliers

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

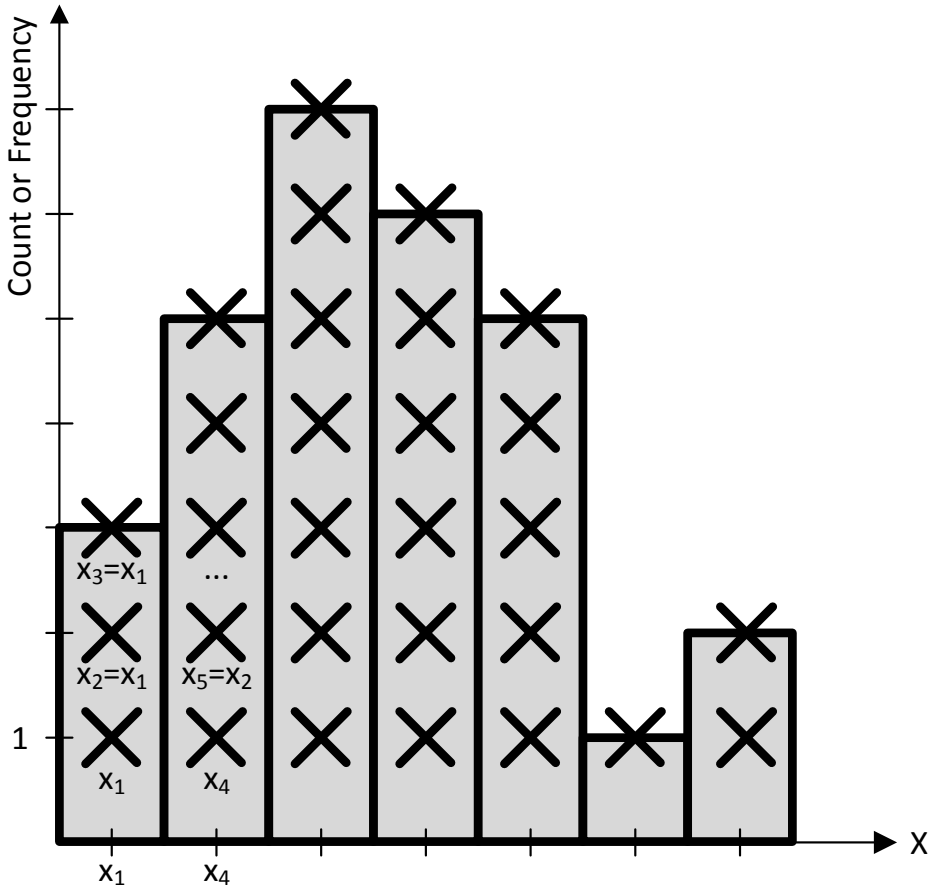
- Outliers are values that are “far” from the central tendency
- No formal definition among statisticians on how to define outliers (how do you define “far”?)
- However, general agreement that they be identified and dealt with appropriately (e.g., keep or discard)
  - They might be the most important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




# Histograms

# Histograms



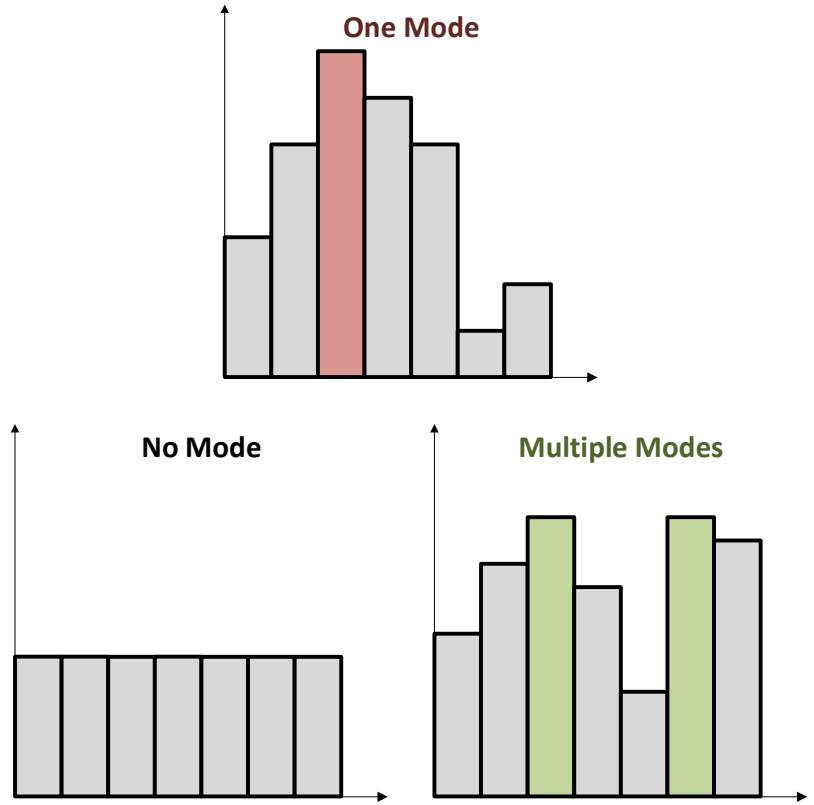

# Codealong: Part D Histograms


# Mode

# Modes and Histograms

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

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




# Codealong: Part E

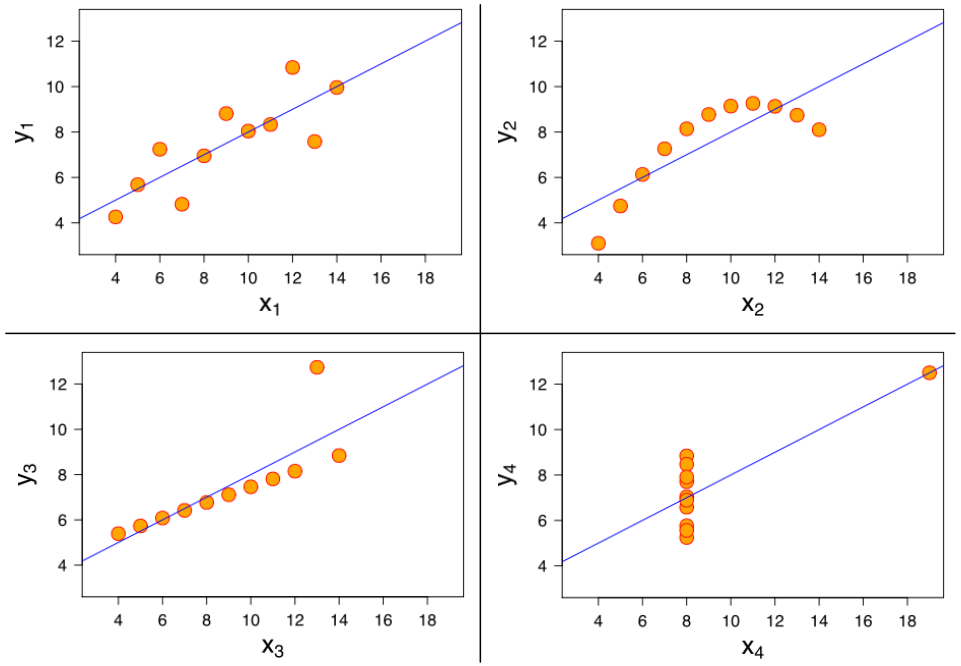
## `.mode()`


# 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.

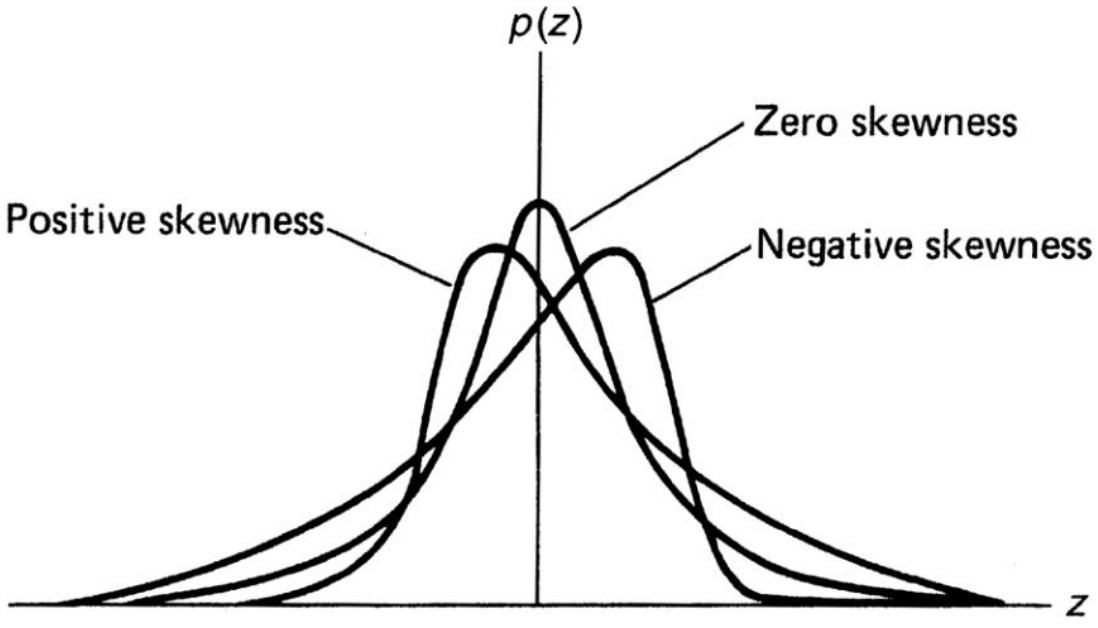
| Property | Value |
|---|---|
| Mean of $x_i$ | 9 |
| Sample variance of $x_i$ | 11 |
| Mean of $y_i$ | 7.50 |
| Sample variance of $y_i$ | 4.122 or 4.127 |
| Correlation between $x_i$ and $y_i$ | 0.816 |
| Linear regression line in each case | $y_i = 3.00 + 0.500\ x_i$ |


# Third and Fourth Moments

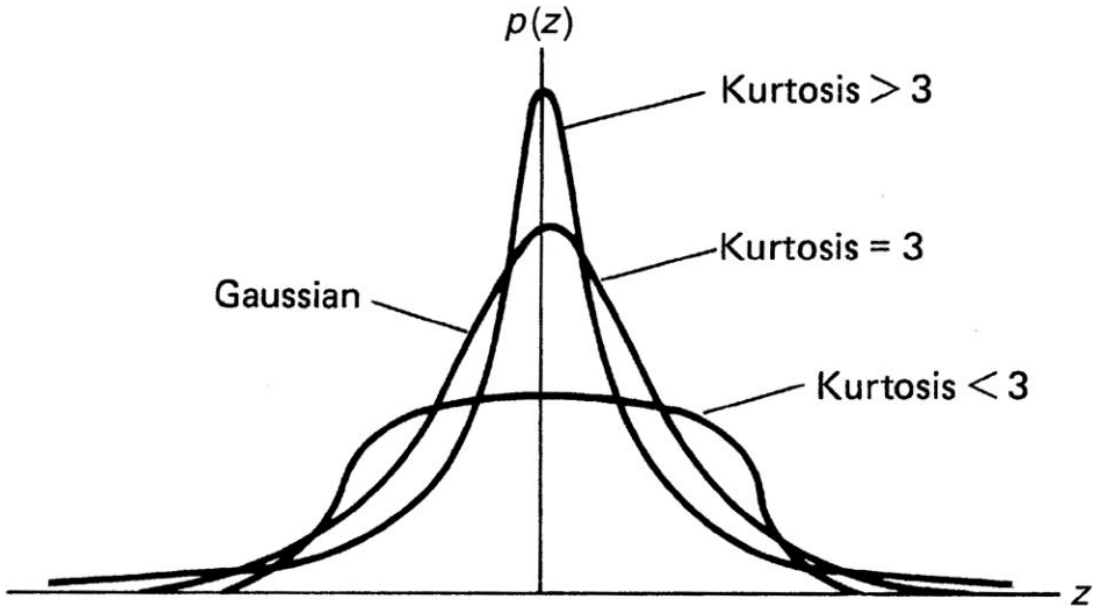
# Skewness

- Skewness measure lack of symmetry. A dataset is symmetric if it looks the same to the left and right of the center point
- a.k.a., the third moment

$$Skew[X] = E[(X - \mu)^3]$$



# Kurtosis



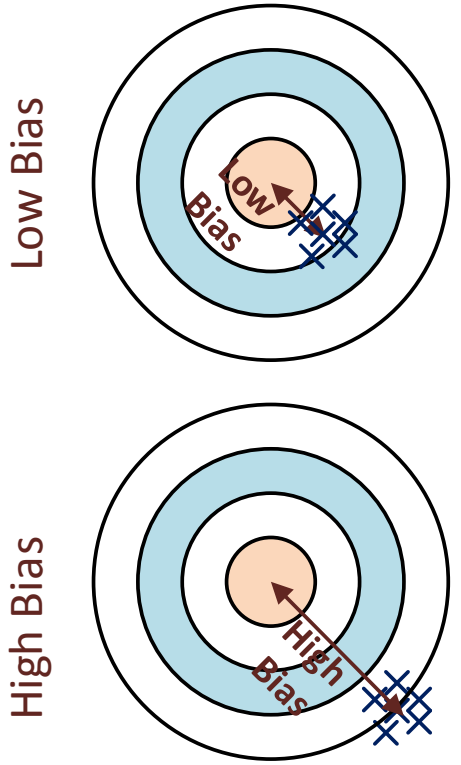

- Kurtosis measures whether the dataset is heavy-tailed (high kurtosis) or light-tailed (low kurtosis) relative to a normal distribution
- Heavy tails signals the presence of outliers
- Light tails the absence of outliers
- a.k.a., the fourth moment

$$Kurt[X] = E[(X - \mu)^4]$$

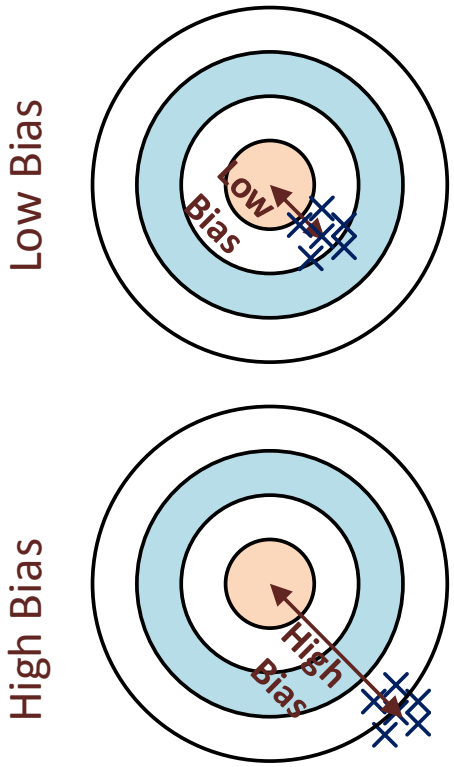

# Measurement Errors

# Bias

- Source of *systematic* rather than *random* error
- Can lead to false conclusion despite the application of correct statistical procedures and techniques



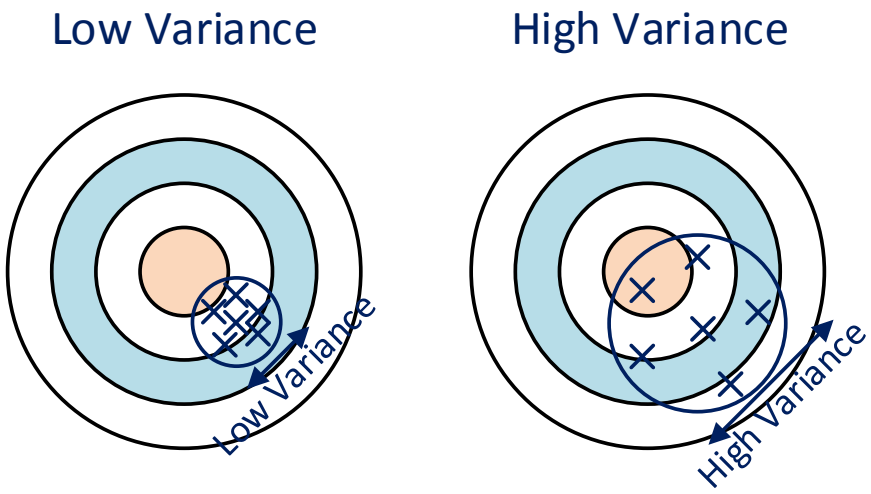
# Bias (cont.)



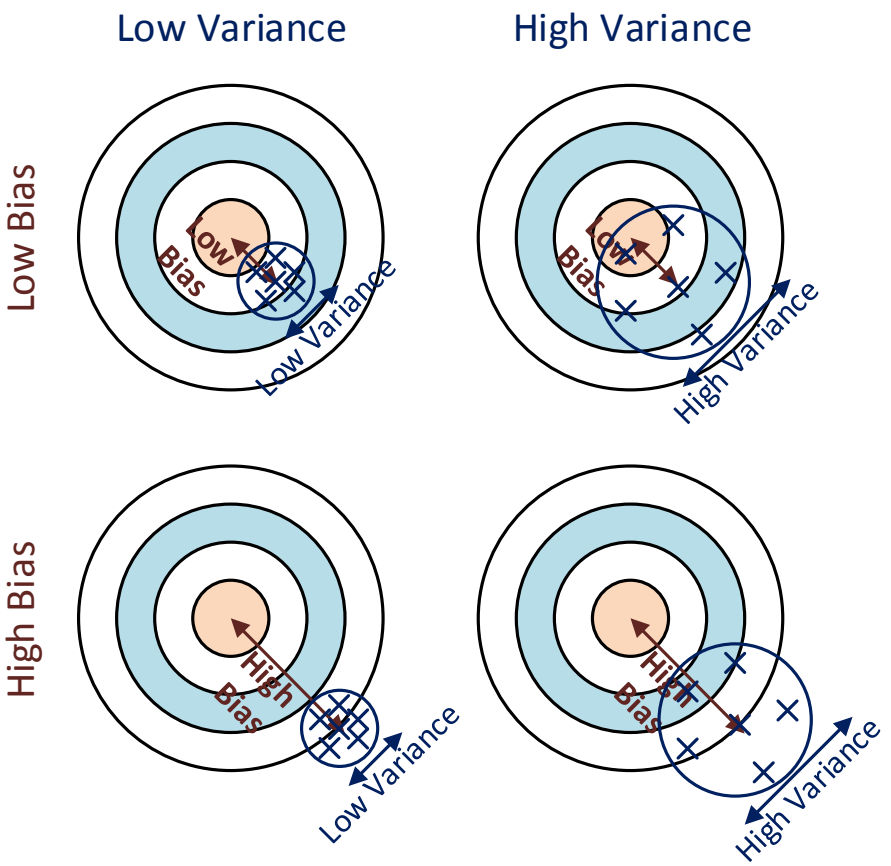

- Selection bias
- Volunteer bias
- Nonresponse bias
- Survival bias

# Variance

- Source of *random* rather than *systematic* error



# Bias vs. Variance, a.k.a, *Systematic* vs. *Random* errors.




# (Linear) Correlation

# Correlation

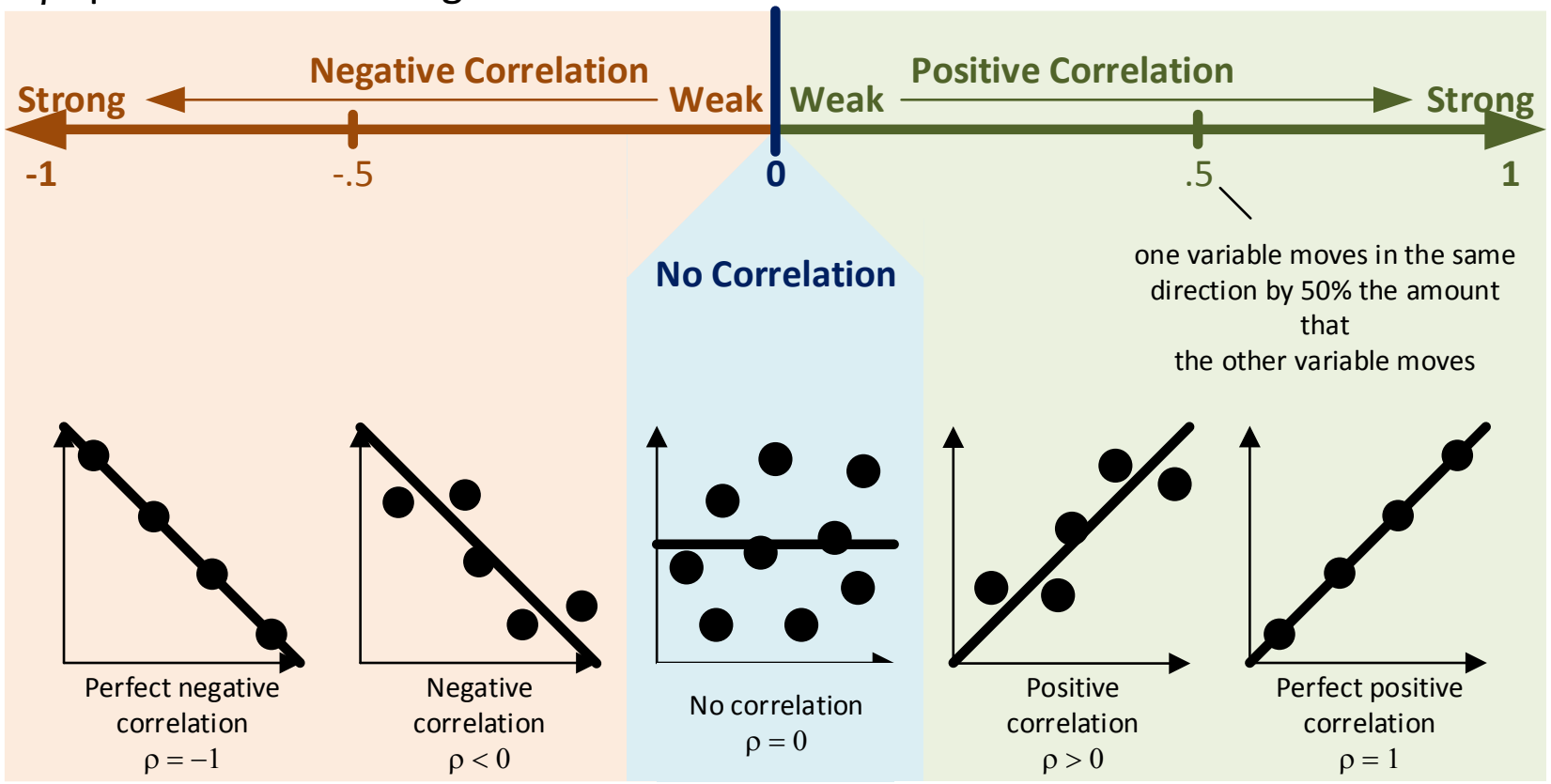
- A measure of strength and direction for a **linear association** between two random variables

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

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

# Correlation (cont.)

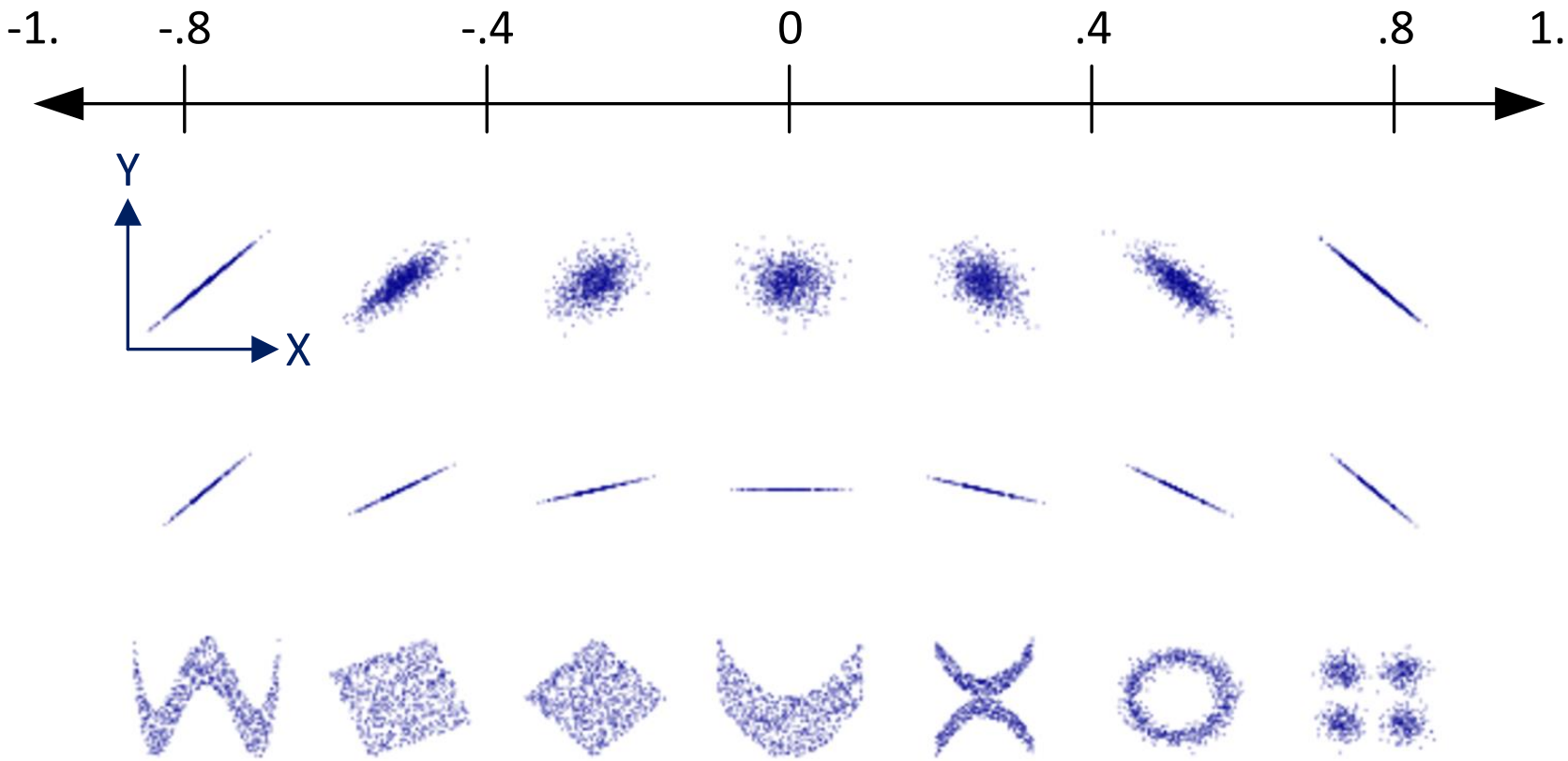
ρ quantifies the strength and direction of movements of two random variables



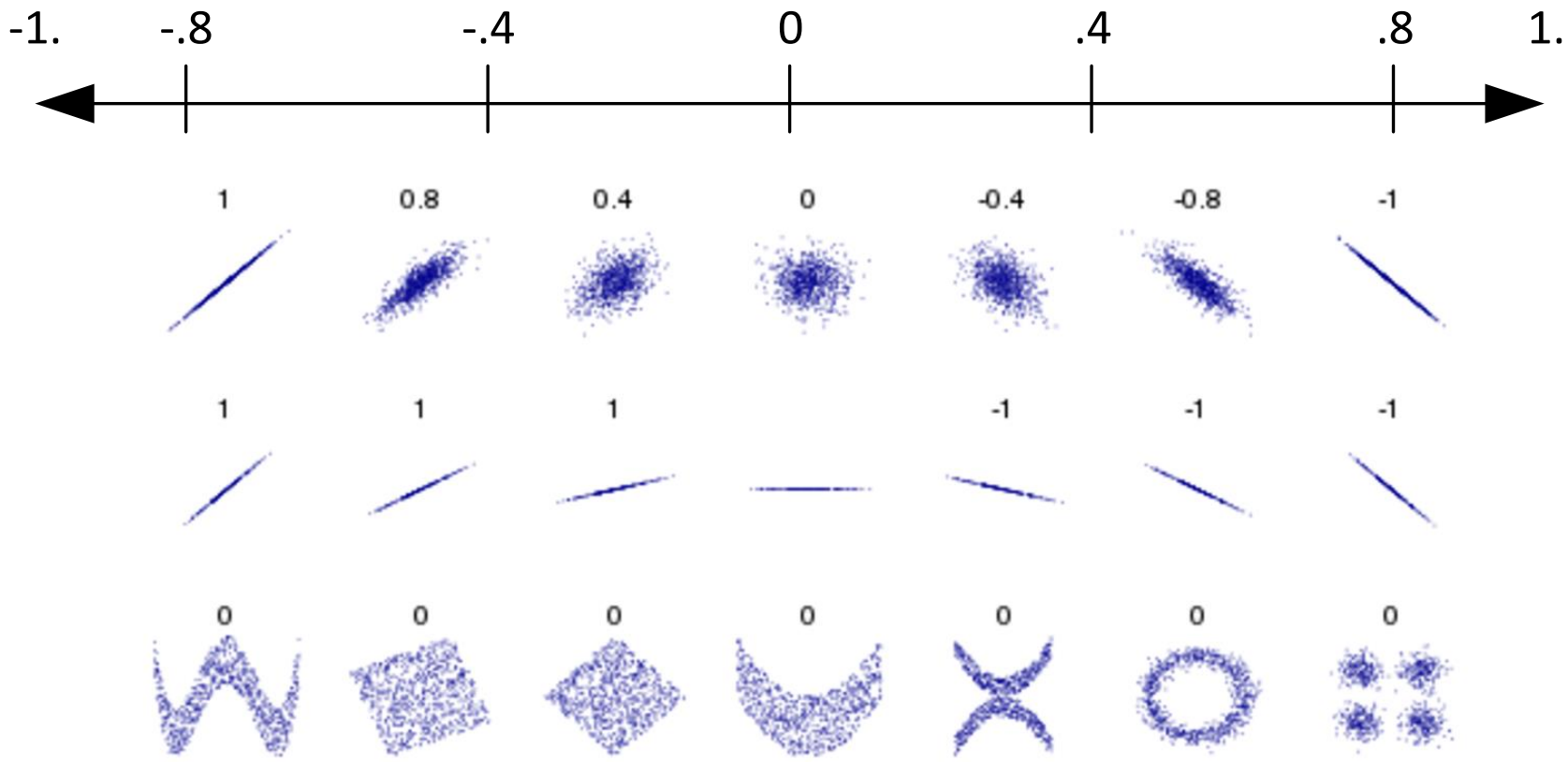

# Activity: Correlations and Scatter Plots

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

EXERCISE



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




# Codealong: Part F

`.corr()`

Heatmaps

Scatter plots

Scatter matrices

# Going further

- *pandas* documentation
  - http://pandas.pydata.org/pandas-docs/stable/


# Lab

# Today's Closing Thought

Forbes / Tech

The Little Black Book of Billionaire Secrets

JAN 15, 2016 @ 06:14 AM

## Microsoft R: One Big Data Tool To Rule Them All?


**Adrian Bridgwater,** CONTRIBUTOR

*I track enterprise software application development & data management.*

FOLLOW ON FORBES (70)

Opinions expressed by Forbes Contributors are their own.


*'Microsoft R Open' — a product name almost worth getting T-shirts printed for, were it not grammatically incorrect. Redmond's big data analytics dream builds one tool to rule them all, maybe... Image: Wikipedia*

Microsoft MSFT +2.77% wants a slice of the big data analytics pie. Truth be told, it has already baked and served itself up a portion by acquiring the R-language and data crunching specialist Revolution Analytics, a purchase it completed in spring of 2015.

In non-developer-speak then, R is a popular open-source statistical computing language well suited to the 'new' world of enterprise class big data analytics. For the record, we used to call this stuff 'data mining' back in the 1990s (some people still do), so don't believe ALL the big data hype you read — regardless, times have changed and we're better at it now.

> In Microsoft's own words, the pitch here is as follows, "Microsoft R Server is your flexible choice for analyzing data at scale, building intelligent apps and discovering valuable insights across your business."

### Four key elements of big data analytics

Named (most probably) after its founders Ross Ihaka and Robert Gentleman, R performs at its best when used for big data statistics, predictive modeling and machine learning. In terms of use, we must now appreciate that there is more than one type of analytics 'thing' or 'function' that we might want to do:

- Big data analytics can be **data preparation** — elements such as de-duplication and time stamping data to know when it was created.
- Big data analytics can be **data exploration** — finding out what the core characteristics of the data set are.
- Big data analytics can be **data visualization** — charts and graphs to make interpreting data trends easier.
- Big data analytics can be **data modeling** — building up the logic so we know how different parts of data relate to each other.

Enough big data foundations already, what Microsoft is doing here is making sure that R Server boasts really good multiplatform support and is essentially open from the core. Remember how Microsoft has (arguably impressively) flummoxed us all by getting the open source religion and preaching it from every minaret in town? This, in effect, is a play for one big data tool to rule them all if you will.

### One tool to rule them all

> In the words of Joseph Sirosh, corporate VP at Microsoft Data Group, "[Microsoft R Server enables] enterprise customers to standardize advanced analytics on one core tool, regardless of whether they are using Hadoop (Hortonworks, Cloudera and MapR), Linux (Red Hat RHT +3.43% and SUSE) or Teradata TDC +1.28%. [We are committed to] building R and Revolution's technology into our broader database, big data and business intelligence offerings and to bring these benefits to customers and students – on-premises, in the Azure cloud and to new platforms."

IDC analyst for business analytics and information management Dan Vesset is convinced that Microsoft is playing an 'important role' (his words) in bringing big data analytics modeling and productivity tools and deployment tools to a broader audience.

> "Advanced and predictive analytics is about developing and testing new models. But it's also about their incorporation by developers into production deployments of decision support and automation solutions that can benefit the whole organization," said Vesset.

### ... and now, it's over to Redmond for the news

The 'news hook' connected to this discussion hinges around the fact that Microsoft has made a new Microsoft R Server Developer Edition (with all the features of the commercial version) now available as a free download — and, the Microsoft Data Science Virtual Machine will include a pre-installed and pre-configured version of Microsoft R Server Developer Edition.

Also, Revolution R Open is now known as 'Microsoft R Open' — a product name almost worth getting T-shirts printed for, were it not grammatically incorrect.

> According to the powers that be in Redmond, "Revolution R Open is now called Microsoft R Open and Microsoft continues its commitment of support for the open source R project, and to releasing regular updates to its enhanced, free distribution of R. Microsoft R Open makes it easier to build reliable applications with R on Windows, Mac and Linux by simplifying the management of R package versions. Microsoft R Open is 100% compatible with all R scripts and packages, and just like R is open source and free to download, use and share."

Is Microsoft doing well in open source big data analytics? Would that there was a highly amusing sarcastic remark to make as an epitaph here... almost none of the trade press were scathing to any degree whatsoever and one even used 'Hooray!' in the headline. It's all about market domination though isn't it? Microsoft is no charity. That's that about as caustic as we can get here.

Interesting times, ooh arr.


# Review

# Review

## You should now be able to:

- ID 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


# Before Next Class

# Projects

| | | | | | |
|---|---|---|---|---|---|
| **Unit Project**<br><br>You will design a research project, perform exploratory data analysis and build a logistic model to determine what factors affect admission the most | *Research Design* ✓ | Exploratory Data Analysis | Logistic Modeling | Executive Summary with Findings | |
| **Final Project**<br><br>Using a dataset of your choosing, you will design a project, build a data science model and present their finding to the course | Lightning Presentation | Experimental Write-up | Exploratory Analysis | Notebook Draft | Final Presentation |

# Unit Project 2 – Exploratory Data Analysis

- Read in your dataset, determine how many samples are present, and ID any missing data
- Create a table of descriptive statistics for each of the variables
  - n, mean, median, standard deviation
- Describe the distributions of your data
- Plot box plots for each variable
- Create a covariance matrix
- Determine any issues or limitations, based on your exploratory analysis

# Unit Project 2 – Exploratory Data Analysis

- Bonus
  - Replace missing values using the median replacement method
  - Log transform data to meet normality requirements
  - Advanced Option
    - Replace missing values using multiple imputation methods


# Exit Ticket

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