Excel Plotting

Data Boot Camp

Lesson 1.3

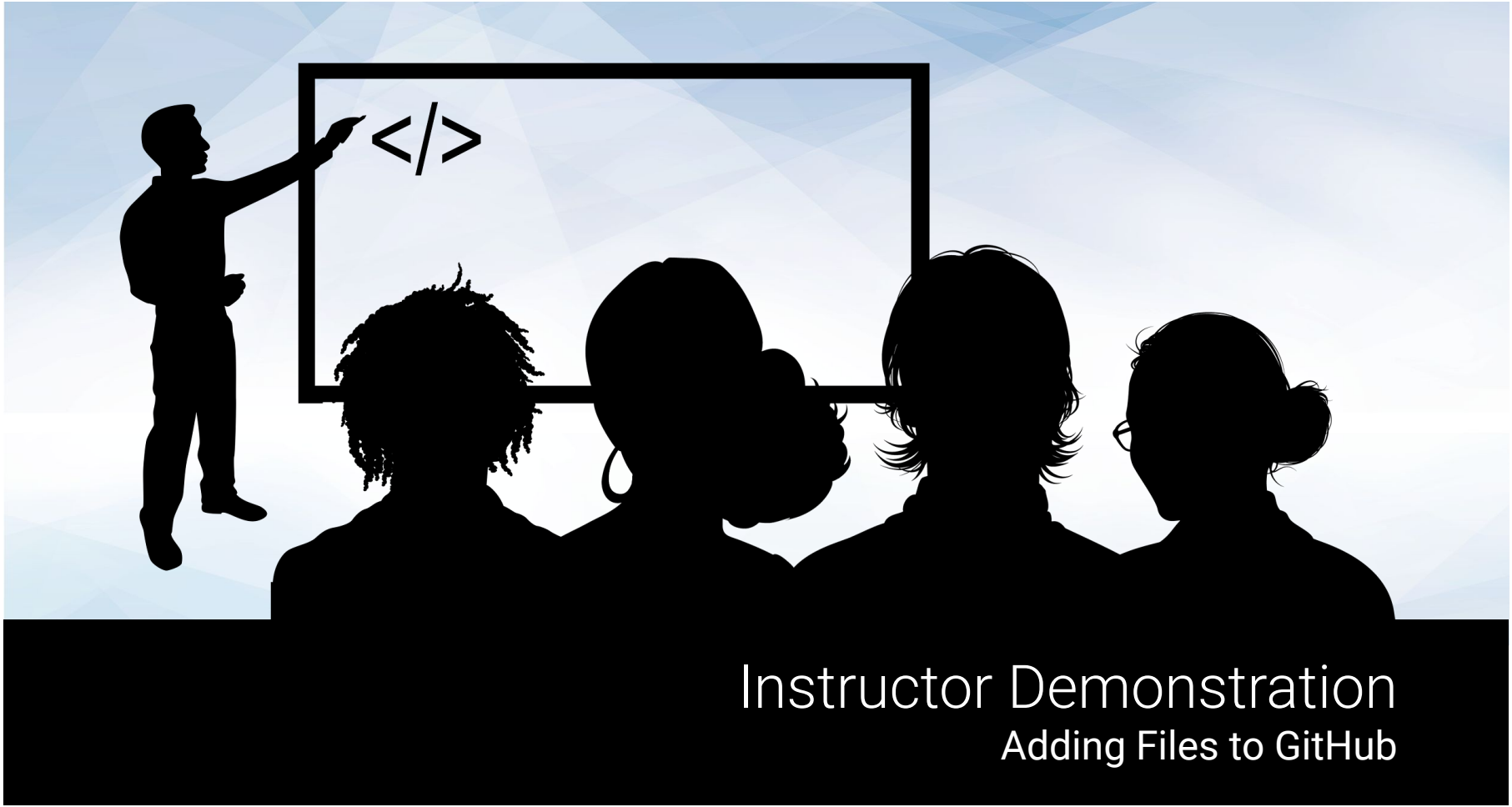
© 2019 Trilogy Education Services, Inc.


WELCOME

We are off to the races!


This will be you
at the end of class.

Let’s introduce our teaching staff to the expanded class!



Instructor Demonstration

Adding Files to GitHub

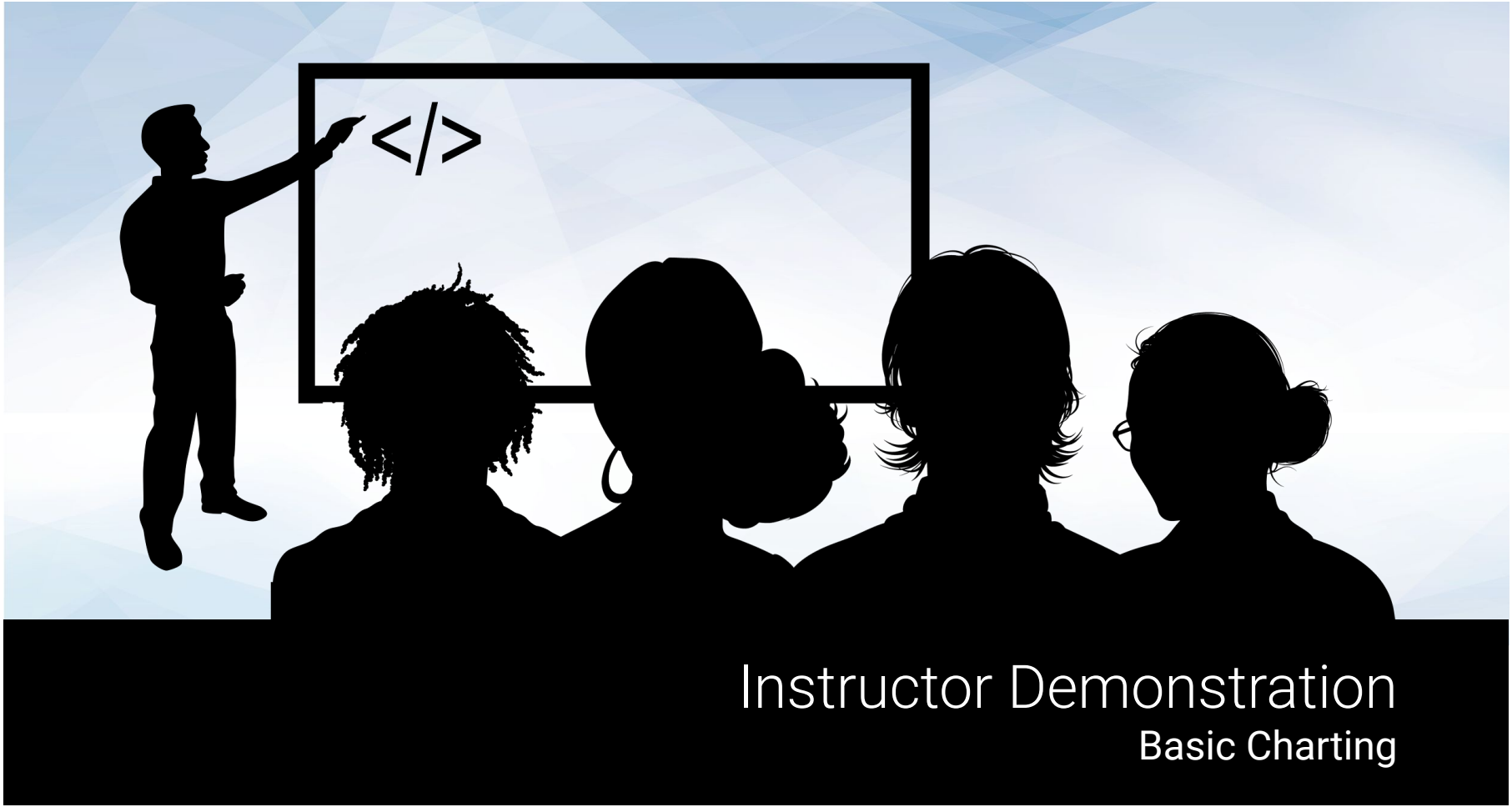
GitHub is a hosting service for source code

- Web interface for **Git**
- **Git** is version-control software
  - Tracks source-code history
  - Allows for collaboration on the same code files across a team or organization
  - Easily updates and rolls back software versions
- Since 2019, GitHub has been used by over 2.1 million companies
- Proficiency in Git and GitHub are highly desirable skills in many industries

We will use Git and GitHub throughout the curriculum

- You will submit your homework assignments using GitHub
- Your individual project work will be version-controlled using Git
- You will be collaborating with teammates using GitHub
- By the end of the curriculum, you should be proficient with the basic Git and GitHub uses

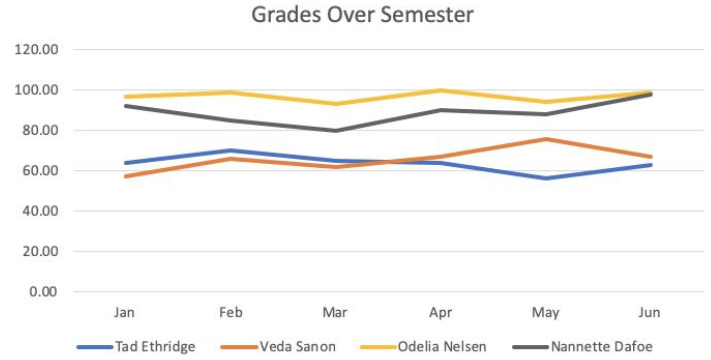
< Demo Time >

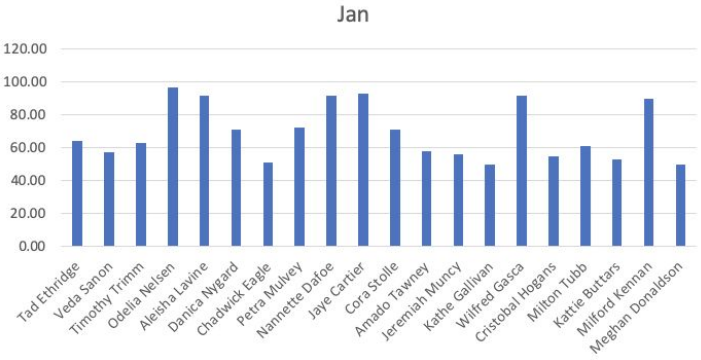


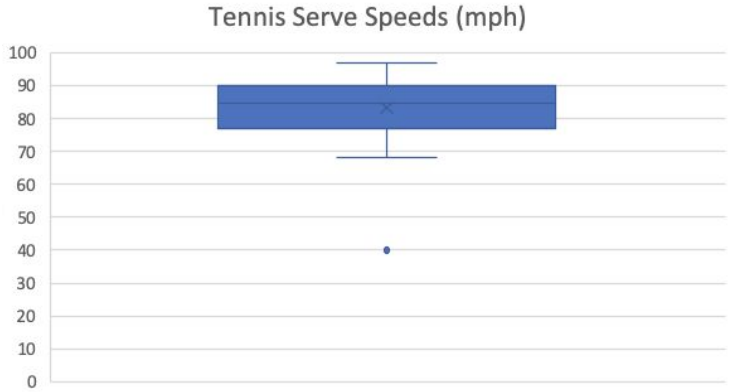
Instructor Demonstration

Basic Charting

It is time to learn Excel visualizations!

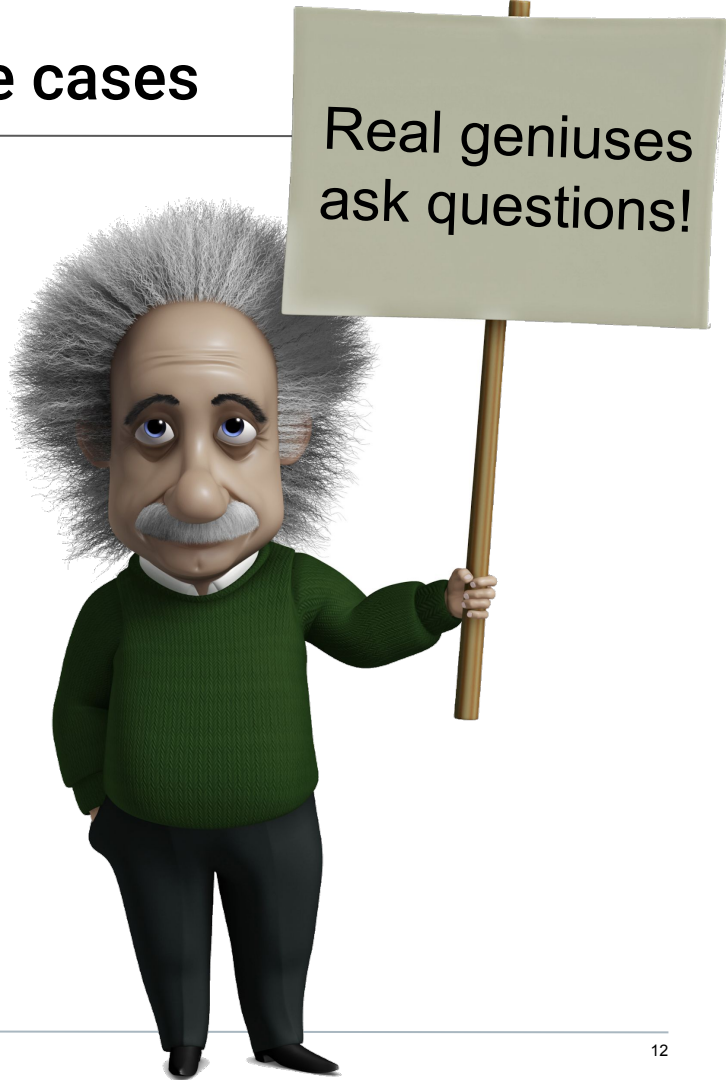




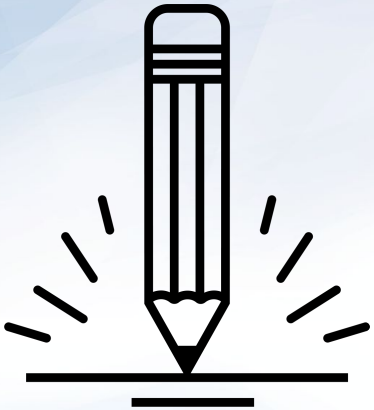


We will look at a few examples and use cases

- Try to follow along!
- In this activity, we will:
  - Go over an example data set
  - Select data of interest
  - Visualize selected data
  - Add labels and titles to our visualization
- Do not hesitate to ask questions
- Our TAs will slack out images for each operating system



< Demo Time >



Activity: The Line and Bar Grades

For this activity, you'll take on the role of the teacher as you create bar and line graphs to visualize your class's grades over a semester.

Suggested Time:
15 Minutes

Activity: Line and Bar Grades

For this activity, you'll take on the role of the teacher as you create bar and line graphs to visualize your class's grades over a semester.

Instructions:

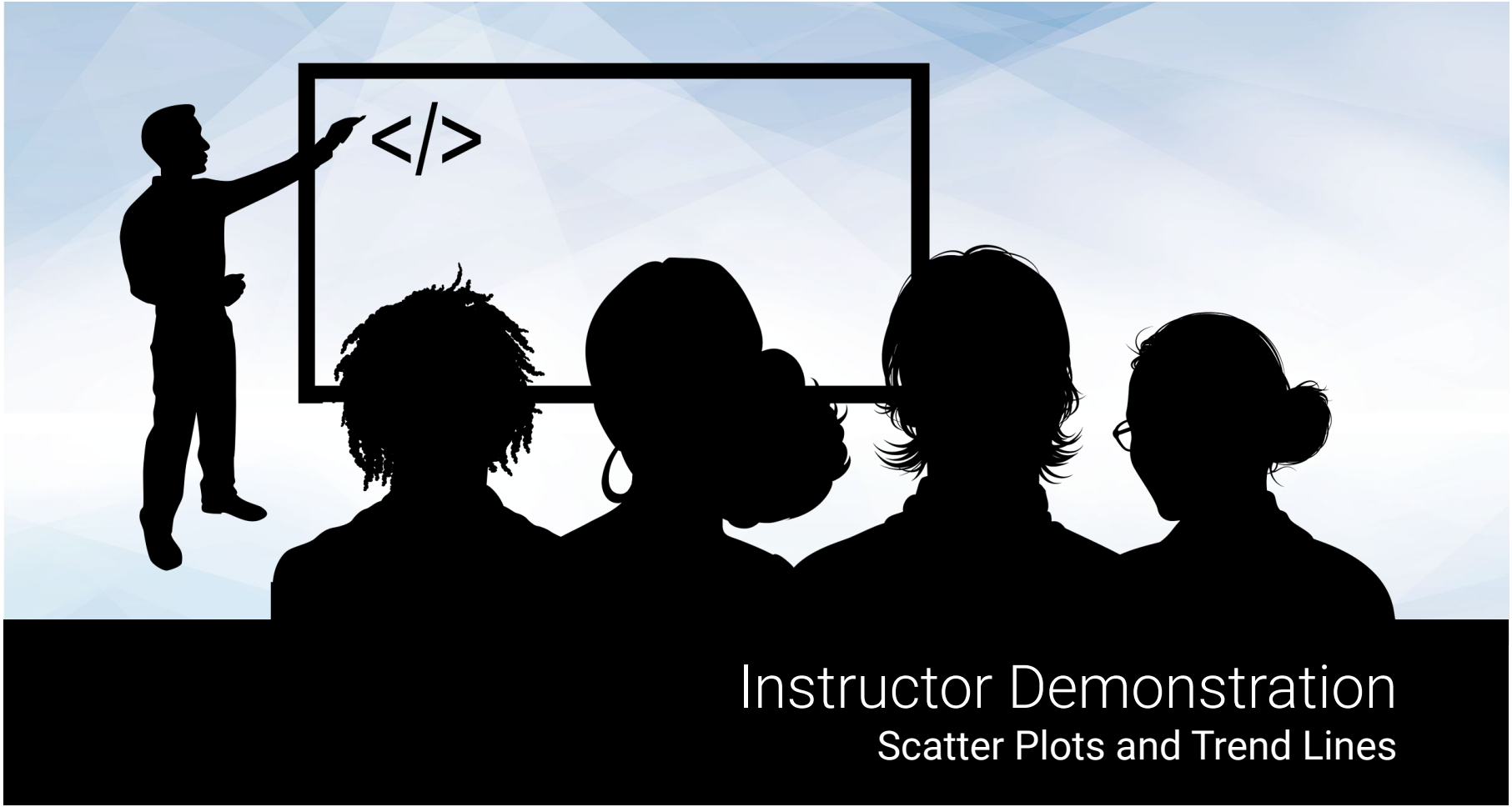
- Create a series of bar graphs that visualize the grades of all students in the class, one graph for every month.
- Create a line graph using all available data to compare students' grades across the semester.
    - Use filtering in the line graph to drill down to a specific student's progress throughout the semester.

Hint:

- When duplicating bar graphs, it helps to get the formatting and look of the chart where you want it for the first graph (that is, for January). Then, copy that chart, and re-select the data for the each subsequent copy.
    - So, keep the style and format, but change the data.

Suggested Time: 15 minutes

Time’s Up! Let’s Review.



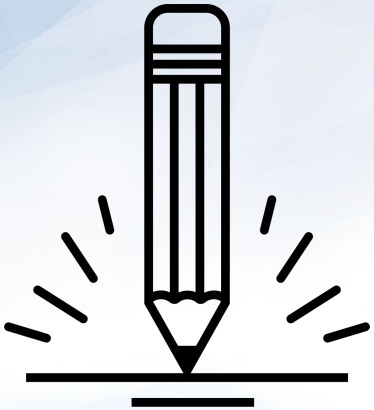
Instructor Demonstration

Scatter Plots and Trend Lines

The scatter plots is a powerful visualization tool!

- Visualizes the comparison between two variables
  - **One variable** is located on the *x*-axis
  - **Another variable** is plotted on the *y*-axis
  - Each data point represents a pair of measurements
- Measurements on a scatter plot are **independent**
- Scatter plots can help to identify positive or negative relationships between two variables
  - Adding a trend line to a scatterplot can visualize this relationship even easier!

< Demo Time >



Partner Activity: Home Sales

For this activity, you will work in pairs to create a series of scatter plots that compare home prices in the St. Louis, MO, region.

Suggested Time:
15 Minutes

Partner Activity: Home Sales

- Create a scatter plot that compares the price of the home with the square feet of the home (`sqft_living`). Make sure to add in axis titles, a chart title, and a trend line.

- Create a scatter plot that compares the price of the home with the number of bedrooms. Make sure to add in axis titles, a chart title, and a trend line.

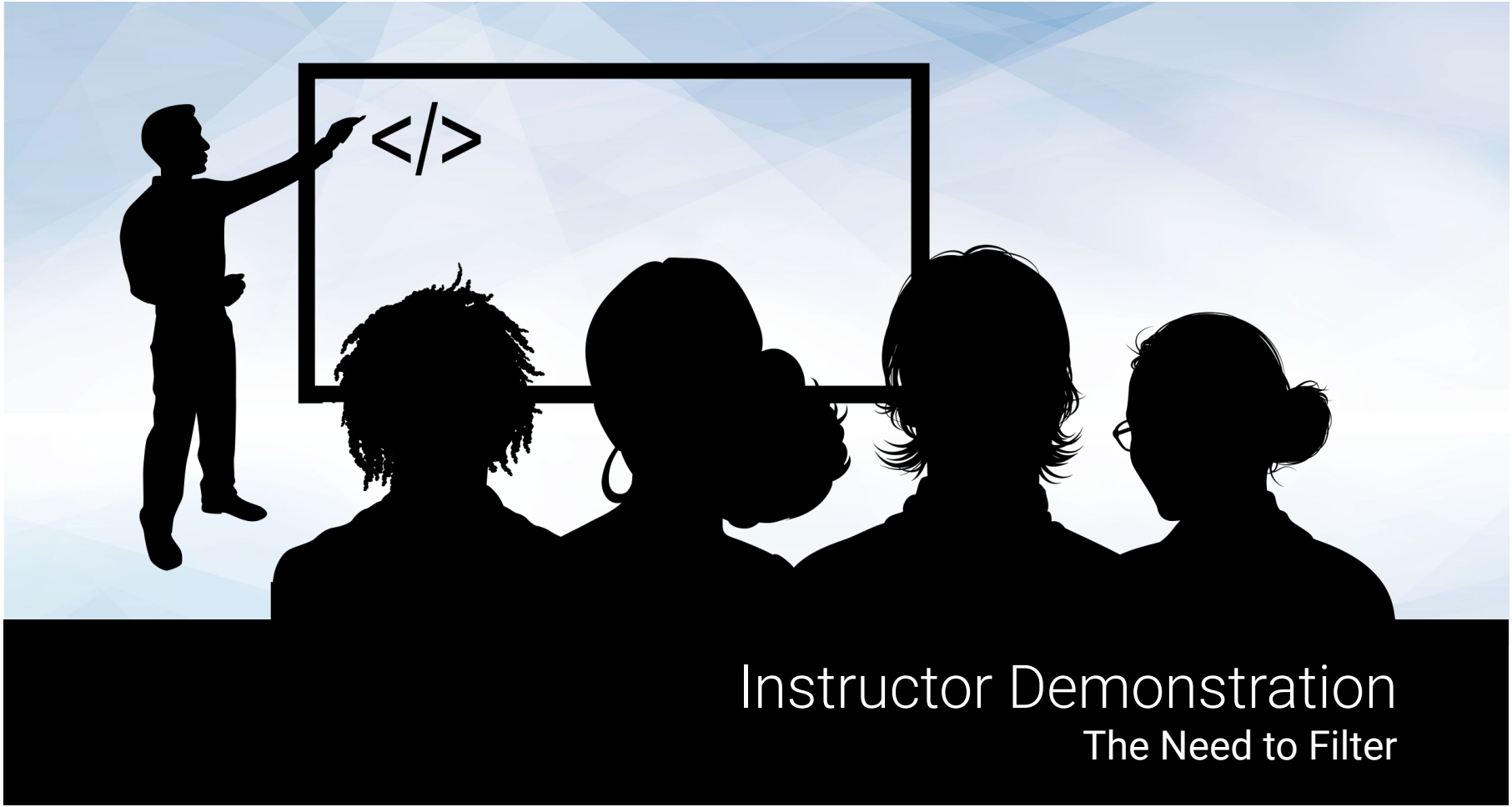
- Create a scatter plot that compares the price of the home with the number of bathrooms. Make sure to add in axis titles, a chart title, and a trend line.

- Go back into each of your charts, and modify the value range on each axis so that they are consistent across charts.

    - We want the axes to match so the data is conveyed in a consistent, truthful manner.

Suggested Time: 15 minutes

Time’s Up! Let’s Review.



Instructor Demonstration

The Need to Filter

Did you notice anything about the data from the last activity?

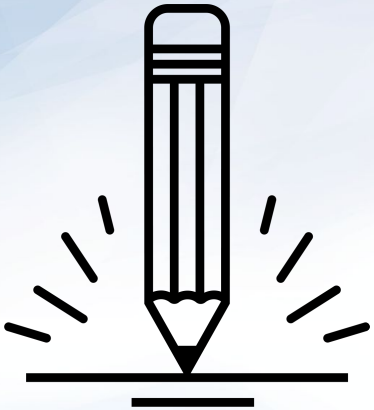
id	date	date_built	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
500652b9-4377-43fd-946c-62e5e8b74e2a	9/9/19	2010	516160	4	6	3585	13648	2	0	8
96b3b4fb-0e10-4ae7-ae31-f259d8971eb5	6/9/21	2014	401820	4	3	2445	4054	1	0	82
5b0a962a-1c25-41fd-8928-21bf0830e0a3	7/29/20	2013	196831	2	2	1656	6879	1	0	39
5647b7e8-443d-4517-bef8-7f4ca8a26191	9/12/21	2012	137291	2	3	2356	6922	1	0	70

There was a LOT of unused data

id	date	date_built	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
500652b9-4377-43fd-946c-62e5e8b74e2a	9/9/19	2010	516160	4	6	3585	13648	2	0	8
96b3b4fb-0e10-4ae7-ae31-f259d8971eb5	6/9/21	2014	401820	4	3	2445	4054	1	0	82
5b0a962a-1c25-41fd-8928-21bf0830e0a3	7/29/20	2013	196831	2	2	1656	6879	1	0	39
5647b7e8-443d-4517-bef8-7f4ca8a26191	9/12/21	2012	137291	2	3	2356	6922	1	0	70

- Most data sets contain multiple variables and factors
- It can be difficult to determine what data is useful when exploring a data set
- It can be hard to locate data of interest
- We need to filter our data

< Demo Time >



Partner Activity: Filtering Home Sales

For this activity, you'll create a filtered chart that visualizes increases in waterfront properties over time in the St. Louis Area.

Suggested Time:
15 Minutes

Partner Activity: Filtering Home Sales

In this activity, you will pair up with one of your classmates in order to create a filtered chart that visualizes increases in waterfront properties over time in the St. Louis Area.

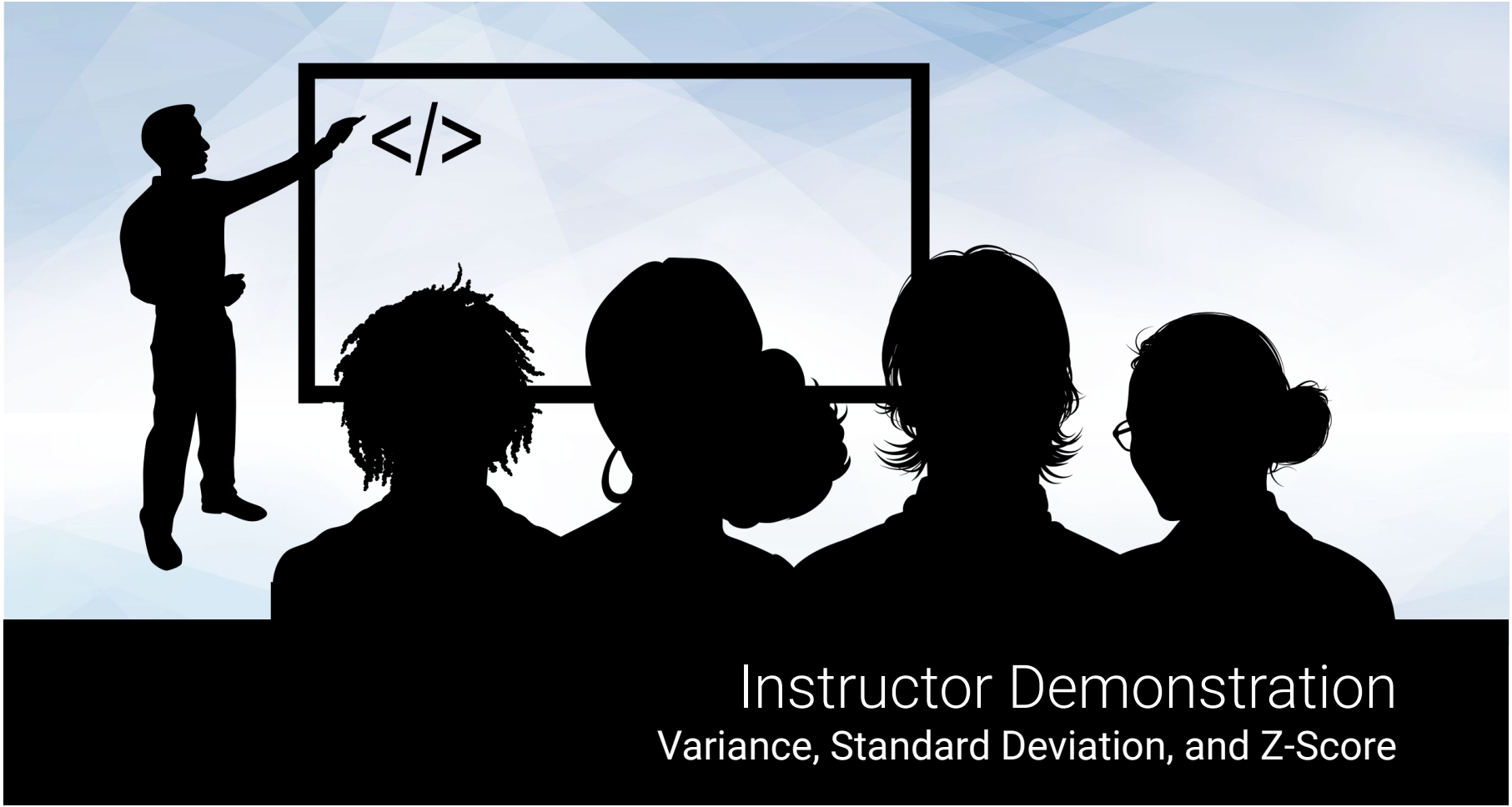
Instructions:

- Use the St. Louis Home Sales Dataset provided.
- Examine the data and check out the available columns.
- Create a line graph that shows the price trend of waterfront homes in St. Louis by the age of the home.

Suggested Time: 15 minutes

Time’s Up! Let’s Review.

Take a Break!



Instructor Demonstration

Variance, Standard Deviation, and Z-Score

Quick Refresher

What are the three measures of central tendency?

The mean, median, and mode.

How are measures of central tendency used?

They are used to describe the center of a data set.

How do you describe the variability of a data set?

Three summary statistical metrics for describing variability

01 Variance

02 Standard Deviation

03 Z-Score

Variance

- Variance is used to describe how far values in the data set are from the mean
    - It describes how much variation exists in the data
- Variance considers the distance of each value in the data set from the center of the data

- σ^2 - the variance
- Σ - sum of all values on the equation line
- μ - the mean of the data set
- N - the number of data points

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

<Time to calculate variance>

Standard Deviation

- Standard deviation describes how *spread out* the data is from the mean
- It is calculated from the square root of the variance
- It uses the same units of measurement as the mean

- σ - standard deviation
- σ^2 - the variance

$$\sigma = \sqrt{\sigma^2}$$

<Time to calculate standard deviation>

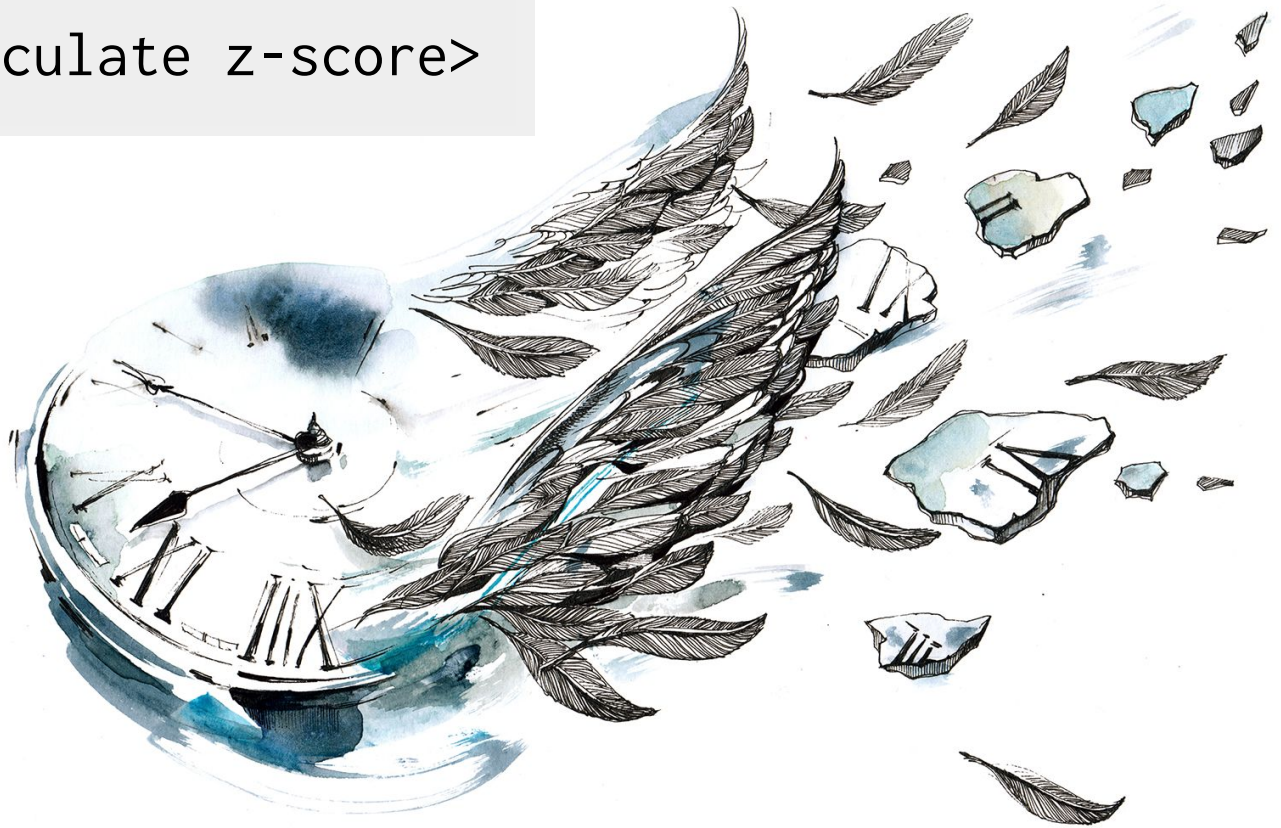
Z-Score

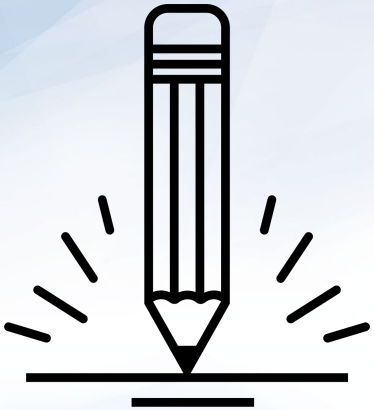
- Z-score describes a single value's distance from the mean of the data set
  - The distance is measured in terms of standard deviations
- Z-score can be positive or negative
  - If negative, the value is less than the mean
  - If positive, the value is greater than the mean
- The smaller the z-score, the closer the value is to the mean

- X - a single value
- μ - the mean of the data set
- σ - the standard deviation of the data set

$$z = \frac{X - \mu}{\sigma}$$

<Time to calculate z-score>





Activity: Variance, Standard Deviation, and Z-Score Review

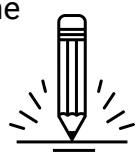
It is now your turn to practice summarizing the variability of a data set using heart disease death rate data from the CDC.

Suggested Time:
15 Minutes

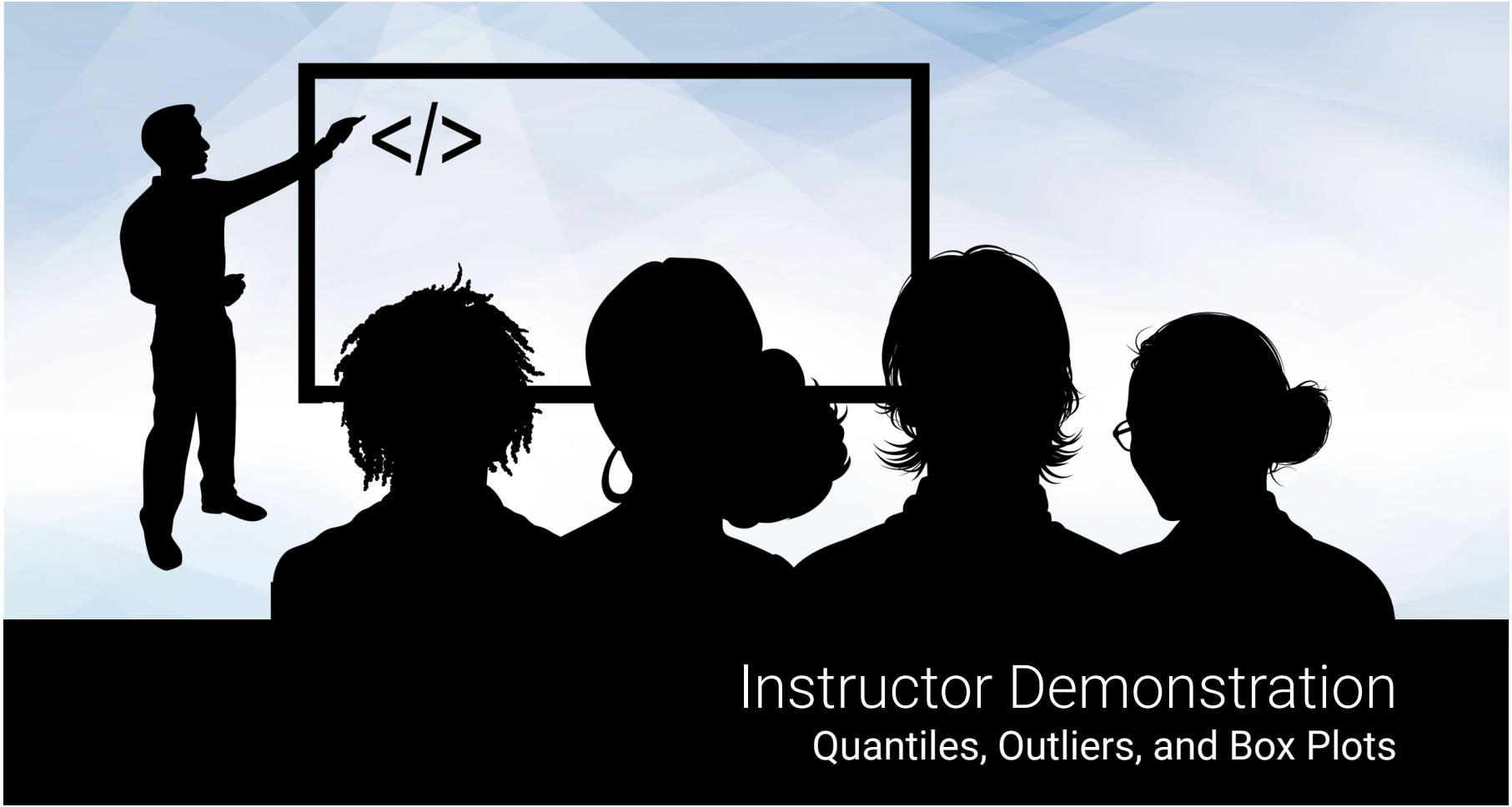
Variance, Standard Deviation, and Z-Score Review Instructions

- Open the variance_review.xlsx workbook that contains your raw data Then clean up the dataset as follows:
  - Rename the **Data_Value** column to **Death Rate Per 100,000**.
  - This column contains missing data, so add a filter to the column that displays all rows except **(blanks)**.
  - Rename the **Stratification1** and **Stratification2** columns to **Gender** and **Race/Ethnicity**, respectively.
  - Rename **LocationAbbr** to **State**.
  - Filter the **GeographicLevel** column so that **State** and **county** values are not compared together.
- Create a new sheet in the workbook named **Summary Table** that has a **State** column containing the following values: **AR** - Arkansas , **CA** - California, **FL** - Florida, **ME** - Maine, **MS** - Mississippi, **OR** - Oregon
- For each state, determine the **mean**, **variance**, and **standard deviation** for the overall death rate.
- Based on your calculated summary statistics determine which state had the greatest difference in death rate across all its counties and which state had the lowest variance in death rate. What was the death rate?
- Create a new sheet in the workbook named **Oregon Z-Scores**. Within this new sheet, copy over the **LocationDesc** (renamed to **County**) and **Death Rate Per 100,000** columns from the raw data for *only* the state **OR** where **Gender** is **Overall**.
- Calculate the **z-score** for the overall death rate by county across the whole state and use those values to determine which county had the largest difference in death rate from the mean of the state.
- Based upon your calculated z-scores, determine which county had the largest difference in death rate from the mean of the state.

Suggested Time: 15 minutes



Time’s Up! Let’s Review.



Instructor Demonstration

Quantiles, Outliers, and Box Plots

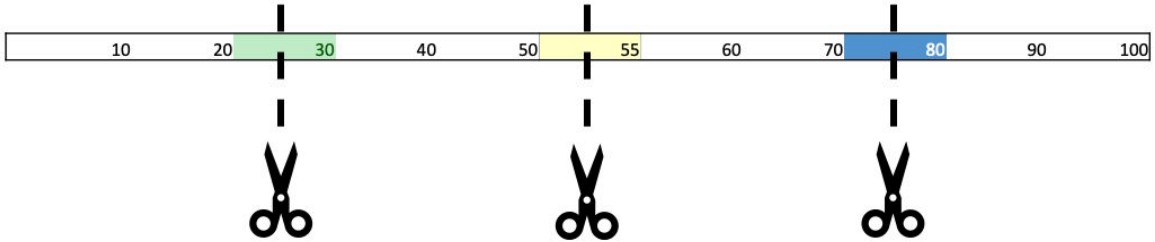
Be careful when describing real-world data

- Real-world data can contain extreme values
- Some summary statistics, such as the mean, take into account *all* values in a data set
- Extreme values can *skew* these statistics!

But how can we summarize real-world data?

We can use quantiles to describe segments of a data set!

- **Quantiles** separate a sorted data set into equal-sized fragments
- Explain that the two most popular types of quantiles are **quartiles** and **percentiles**
  - Quartiles divide the data set into four equal parts
  - Percentiles divide the data set into 100 equal parts



< Demo Time >

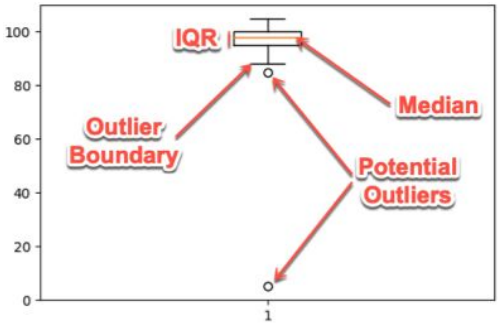
Extreme values may not always be reliable

- In **data science**, extreme values are often suspicious
  - Could the measurement be a mistake?
  - Is the data trustworthy?
- Suspicious values are called **potential outliers**
- An outlier is a data point that differs from the rest of a data set
- Outliers can inaccurately skew a data set
  - Outliers can cause us to misrepresent the actual data

There are two ways to identify potential outliers

01 Qualitatively

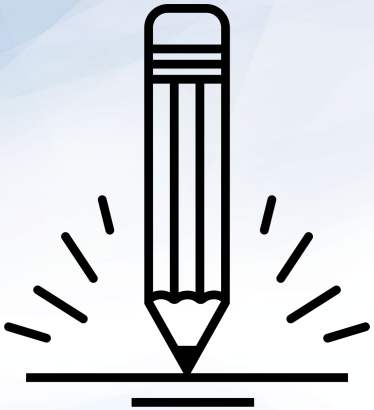
- Use box and whisker plots to visually identify potential outlier data points



02 Quantitatively

- Determine the outlier boundaries in a data set using the “1.5 IQR” rule
  - IQR is the interquartile range, or the range between the 1st and 3rd quartiles
  - Anything **below** Q1 − 1.5 * IQR could be an outlier
  - Anything **above** Q3 + 1.5 * IQR could be an outlier

< Demo Time >



Activity: Outliers - Drawn and Quartiled

In this activity, you will be investigating data from a dataset called 80 Cereals. Your task is to search through the ratings of each product and determine if there are any potential outliers in the dataset.

Suggested Time:
10 Minutes

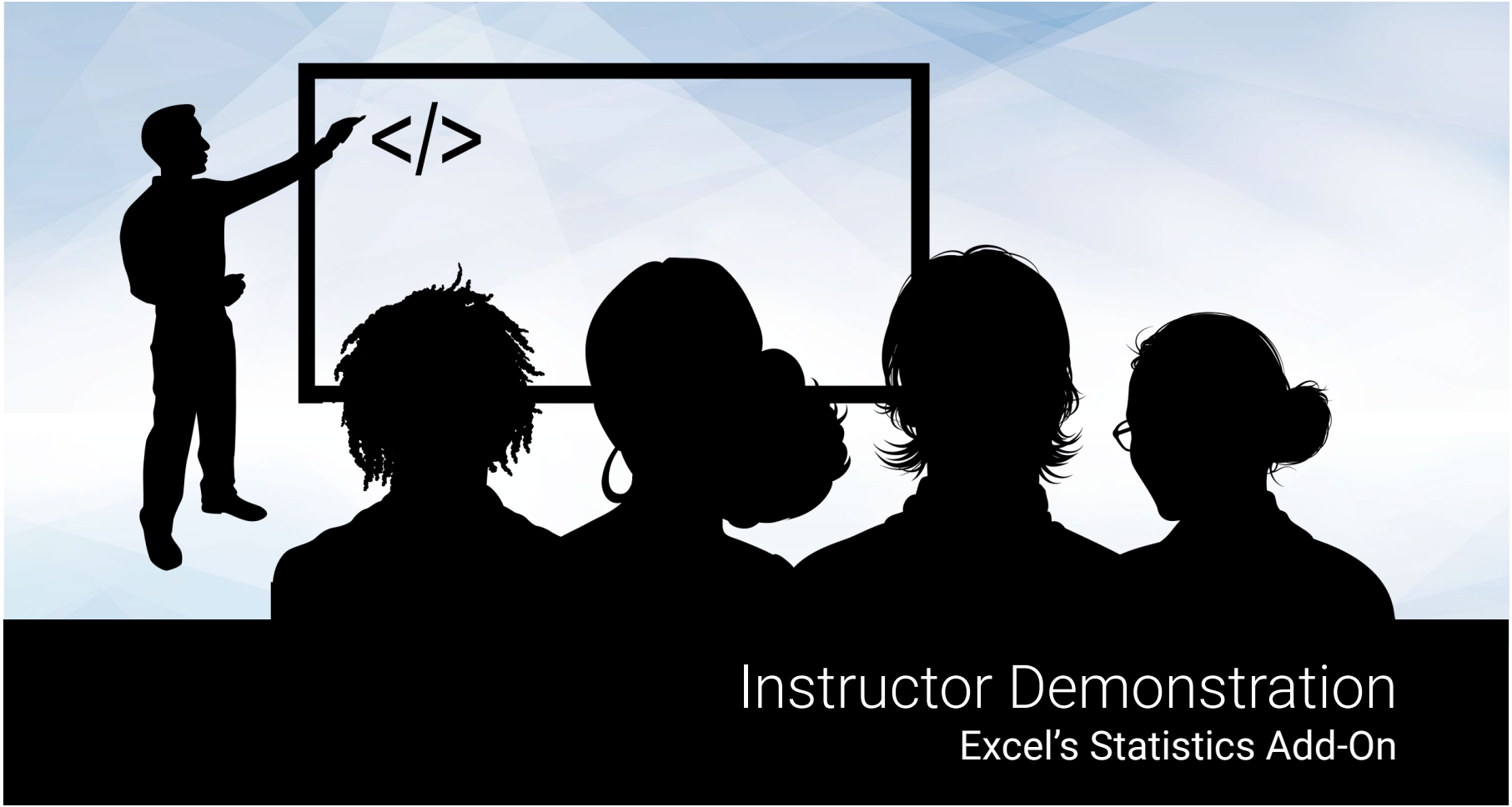
Variance, Standard Deviation, and Z-Score Review Instructions

Instructions:

- Open up the activity workbook, and familiarize yourself with the raw data.
  - File: Unsolved/Outliers_Activity_Unsolved.xlsx
- Create a new worksheet, and name it "Outlier Testing".
- In the "Outlier Testing" worksheet, create a summary statistics table of the Antioxidant_content_in_mmol_100g for the following statistics:
  - Mean
  - Median
  - Minimum value
  - Maximum value
  - First quartile
  - Third quartile
  - Interquartile Range
- Using the calculations from the table, determine the lower and upper boundaries of the 1.5*IQR rule.
- Determine if there are any products whose Antioxidant_content_in_mmol_100g falls outside of the 1.5*IQR boundaries. List those products and their antioxidant content on the worksheet.
- Create a box plot of the Antioxidant_content_in_mmol_100g for all products.
  - **Note**: Be sure to add a title, and label your *y*-axis.

Suggested Time: 15 minutes

Time’s Up! Let’s Review.



Instructor Demonstration

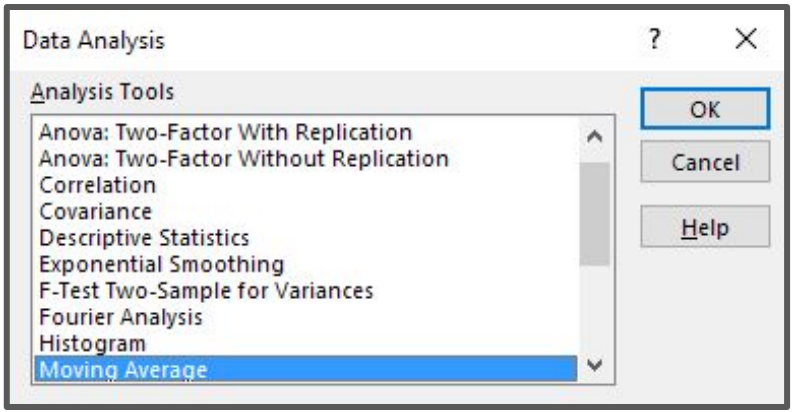
Excel’s Statistics Add-On

Excel is a great foundational tool

Up to this point, we have only covered summary statistics...

But Excel can be used for even **MORE** statistics!

- The Excel Analysis ToolPak contains
  - T-tests
  - Correlation Tests
  - Regression Tests
  - ANOVA
- We will cover all of these functions throughout the course!



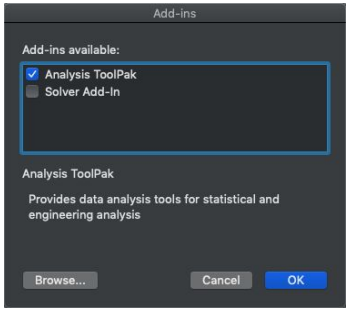
Analysis ToolPak is not designed for in-depth data analytics

- Excel struggles with medium-to-large data sets
  - >200 columns or >100000 rows
  - Depends on machine
- Excel does not automatically record parameters for statistical tests
- Excel's Analysis ToolPak should be used for:
  - Gut checks
  - One-off analysis

How to install and use the Excel Analysis ToolPak (macOS)

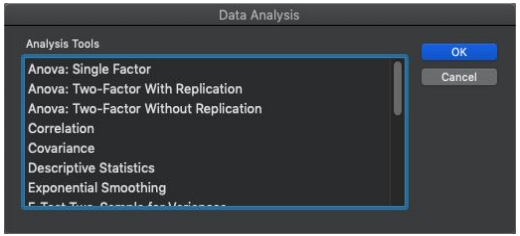
To Install:

1. Go to the “Tools” menu in Excel.
2. Select the “Excel Add-Ins…” option.
3. Enable the “Analysis ToolPak” option.
4. Click “OK.”



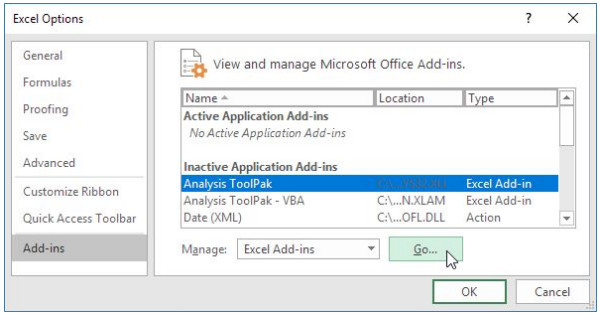
To Use:

1. Go to the “Data” menu in Excel.
2. Select the “Data Analysis” option.



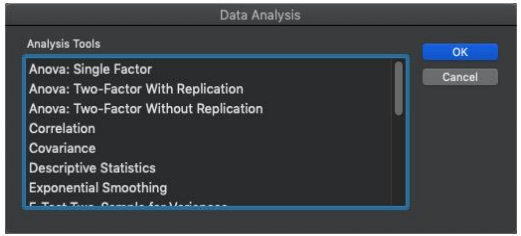
How to install and use the Excel Analysis ToolPak (PC)

To Install:



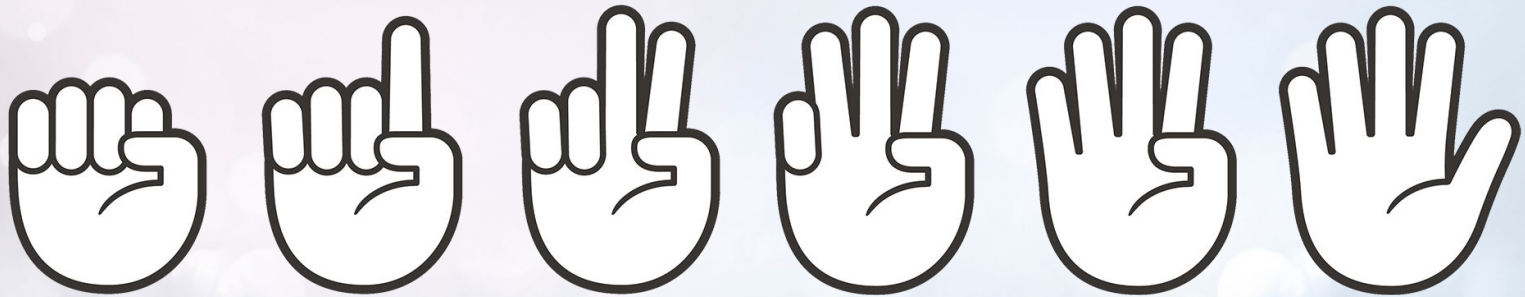
1. Click the “File” tab
2. Go to “Options”
3. Select the “Add-ins” category
4. In the “Manage” box, select “Excel Add-ins” and click Go.
5. In the “Add-ins” box, enable the “Analysis ToolPak” and click OK.

To Use:



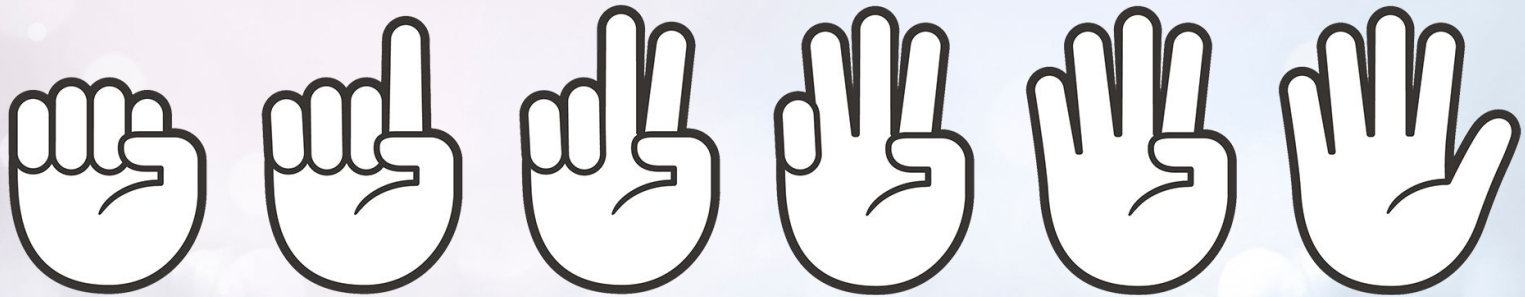
1. Go to the “Data” menu in Excel.
2. Go to the “Analyze” section.
3. Select the “Data Analysis” option.

< Demo Time >



FIST TO FIVE:

Who feels comfortable
with plotting figures in Excel?



FIST TO FIVE:

Who feels comfortable
calculating summary statistics in Excel?