Excel Plotting

Data Boot Camp

Lesson 1.3

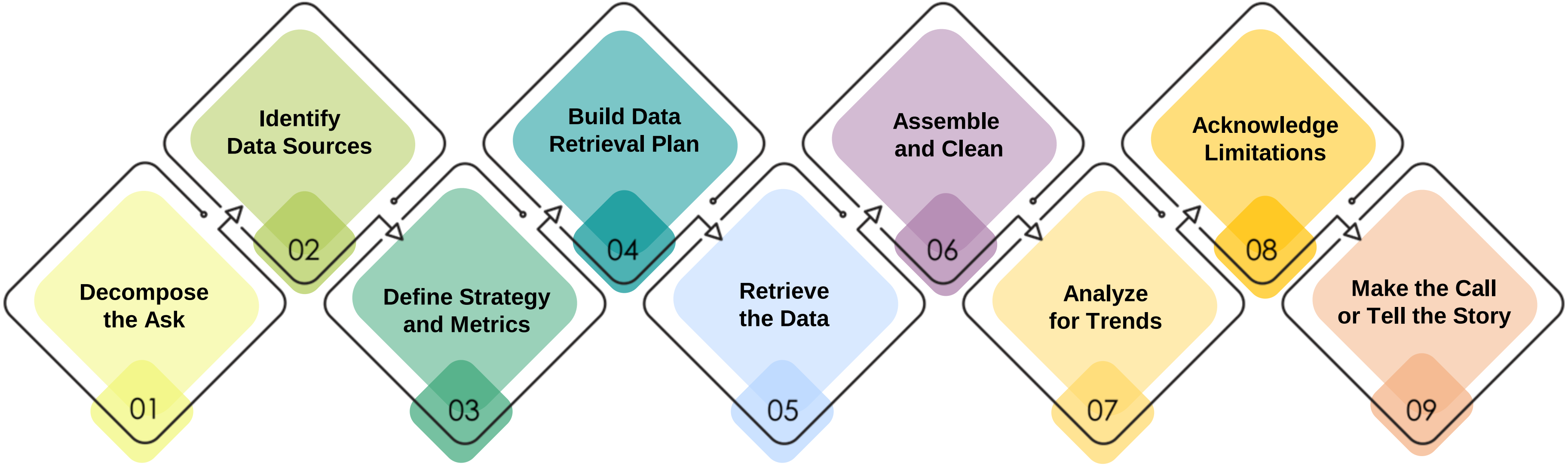
WELCOME

We are off to the races!

This will be you
at the end of class.

Analytics Paradigm

Regardless of type or industry, this paradigm provides a repeatable pathway for effective data problem solving.

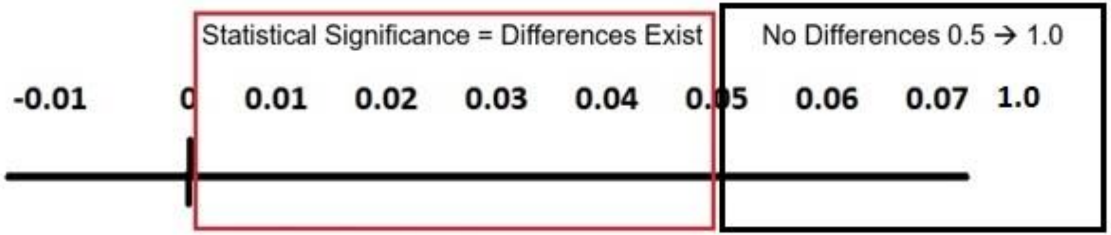


01 Decompose the Ask
02 Identify Data Sources
03 Define Strategy and Metrics
04 Build Data Retrieval Plan
05 Retrieve the Data
06 Assemble and Clean
07 Analyze for Trends
08 Acknowledge Limitations
09 Make the Call or Tell the Story

Alpha is chosen BEFORE
experiment begins

Real Number Line, Calculated p-Values, and the t-Test comparing TWO SAMPLES

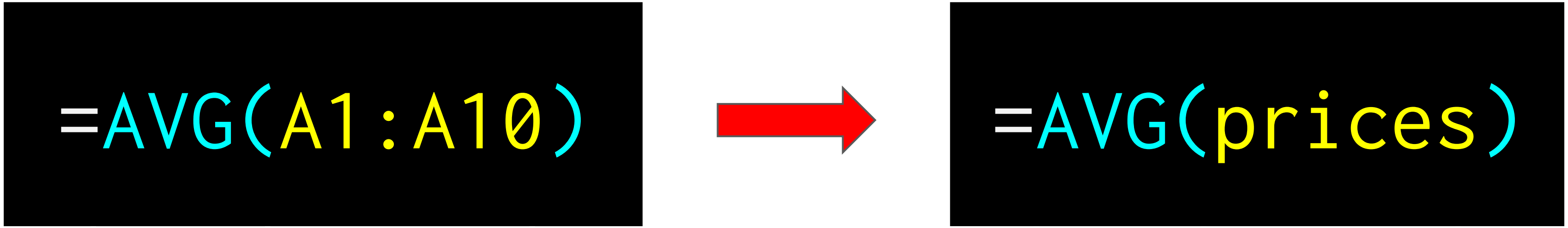
$\alpha = 0.05$



p-Values are calculated AFTER the analysis

There are multiple ways to select data in a formula

But we can name a range of values to make interpreting formulas easier!



```
=AVG(A1:A10)
```

```
=AVG(prices)
```

Ooh…Coding! (Sort Of)

Q But what if we want to **combine** conditions?

A AND, NOT, OR

```
=IF(AND(D2>5, D2<10),TRUE,FALSE)
```

Nested Functions

Three most common measures of central tendency

01 Mean

- The “arithmetic” average
- **To calculate:** The sum of all values, divided by the number of values

02 Median

- The middle value of a data set
- **To calculate:** Sort the data set and find the center

03 Mode

- The most frequent value of a data set
- **To calculate:** Count the frequency of each value in a data set, determine the most frequent value

Formatting in Excel falls into two categories

01 Data Formatting

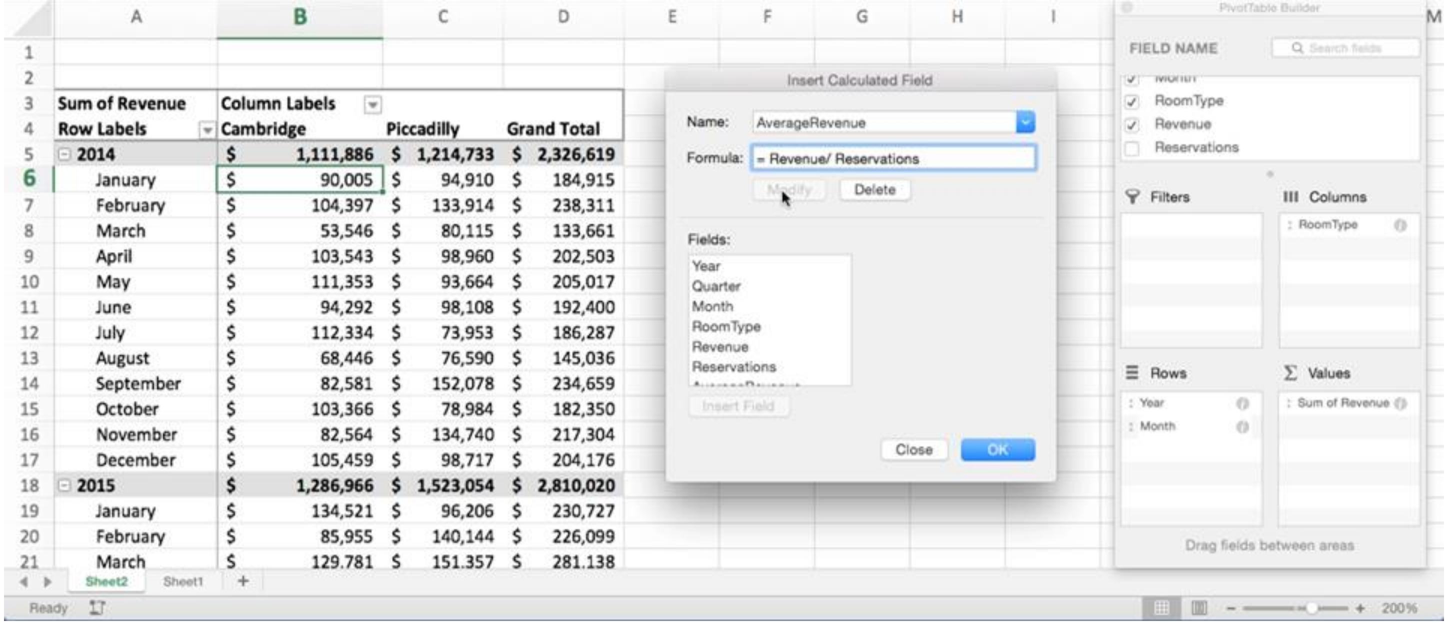
- Changes the way a value is represented in a cell.
- Used to help with interpretation or to add context to the range of values
- Examples:
  - Date and Time
  - Currency
  - Percentage
  - Scientific Notation

02 Style Formatting

- Changes the way the cell and text are viewed
- Can include font color, cell highlighting, borders, etc.
- Can be performed manually or using formulas/logic (conditional formatting)

Get Pivot With It

Pivot tables are one of the most important data visualization concepts to master in this class. (Don’t worry. They are a cinch to deal with.)



Sum of Revenue	Column Labels		
Row Labels	Cambridge	Piccadilly	Grand Total
2014	$ 1,111,886	$ 1,214,733	$ 2,326,619
January	$ 90,005	$ 94,910	$ 184,915
February	$ 104,397	$ 133,914	$ 238,311
March	$ 53,546	$ 80,115	$ 133,661
April	$ 103,543	$ 98,960	$ 202,503
May	$ 111,353	$ 93,664	$ 205,017
June	$ 94,292	$ 98,108	$ 192,400
July	$ 112,334	$ 73,953	$ 186,287
August	$ 68,446	$ 76,590	$ 145,036
September	$ 82,581	$ 152,078	$ 234,659
October	$ 103,366	$ 78,984	$ 182,350
November	$ 82,564	$ 134,740	$ 217,304
December	$ 105,459	$ 98,717	$ 204,176
2015	$ 1,286,966	$ 1,523,054	$ 2,810,020
January	$ 134,521	$ 96,206	$ 230,727
February	$ 85,955	$ 140,144	$ 226,099
March	$ 129,781	$ 151,357	$ 281,138

Insert Calculated Field

Name: AverageRevenue

Formula: = Revenue/ Reservations

Fields:
Year
Quarter
Month
RoomType
Revenue
Reservations

PivotTable Builder

FIELD NAME
☑ RoomType
☑ Revenue
□ Reservations

Columns: RoomType

Rows: Year, Month

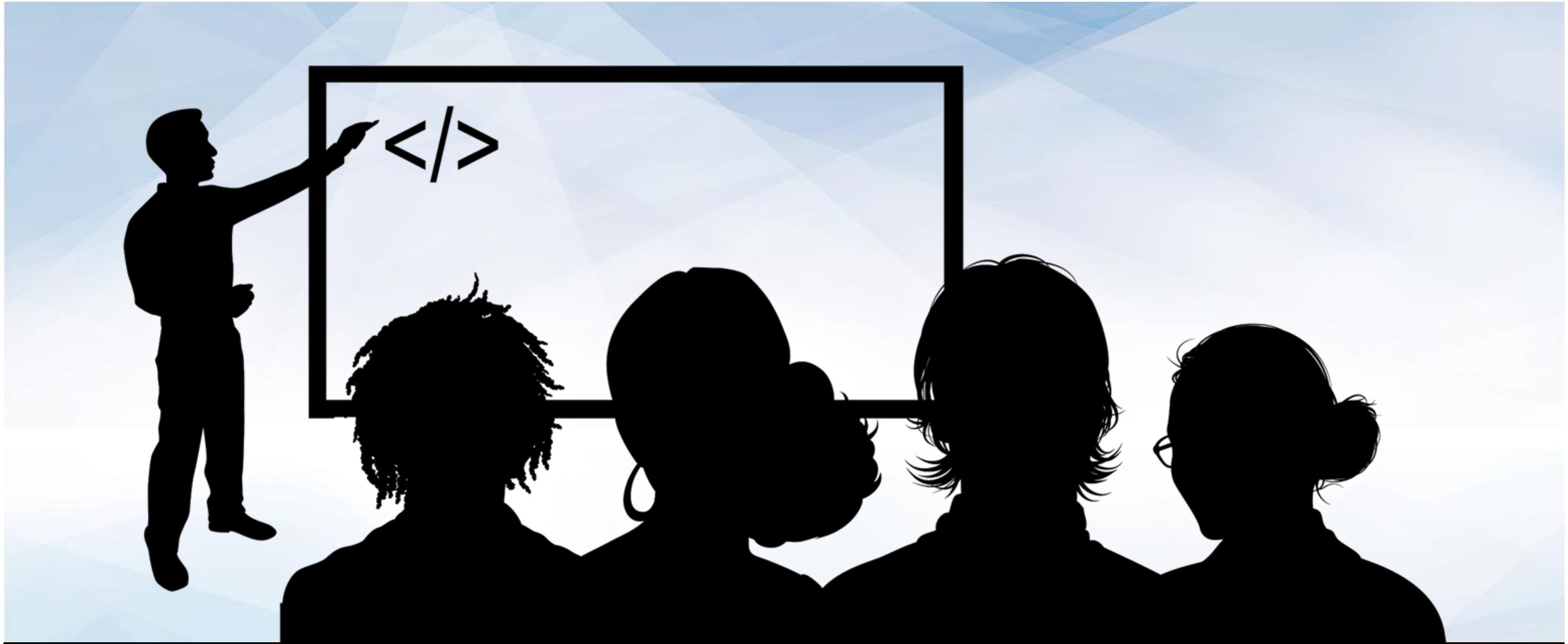
Values: Sum of Revenue

Look It Up with Lookups

What will this yield?

```
=vlookup( “Asteroid 9”, Planets, 3, FALSE)
```

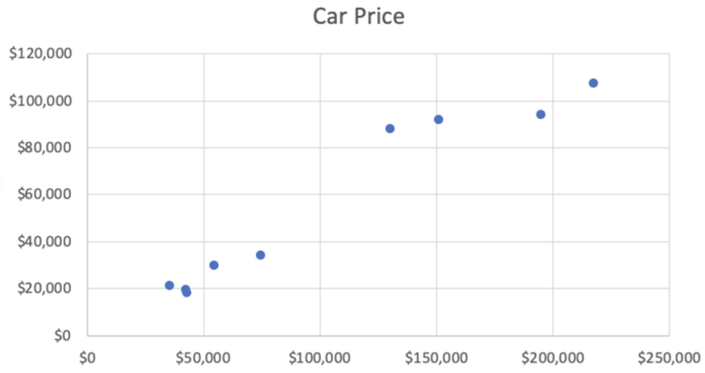
Planet	Population	Species
Zeelo	5020	Zoltans
Merinoa	380	Murphies
Cardboard Box	2	Hambones
…	…	
Asteroid 9	95	Asterisks

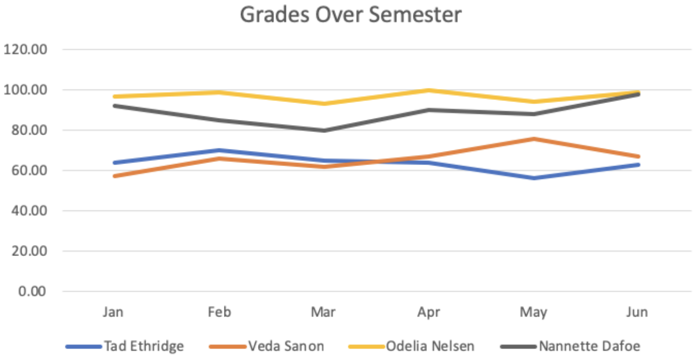


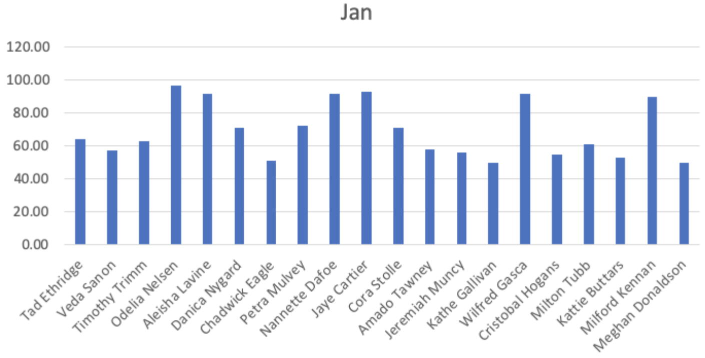
Instructor Demonstration

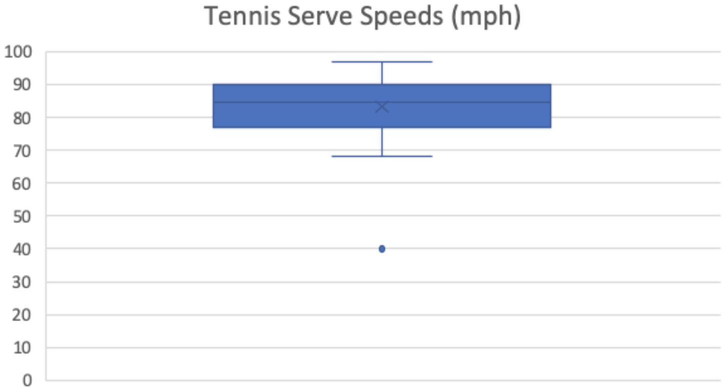
Basic Charting

It is time to learn Excel visualizations!



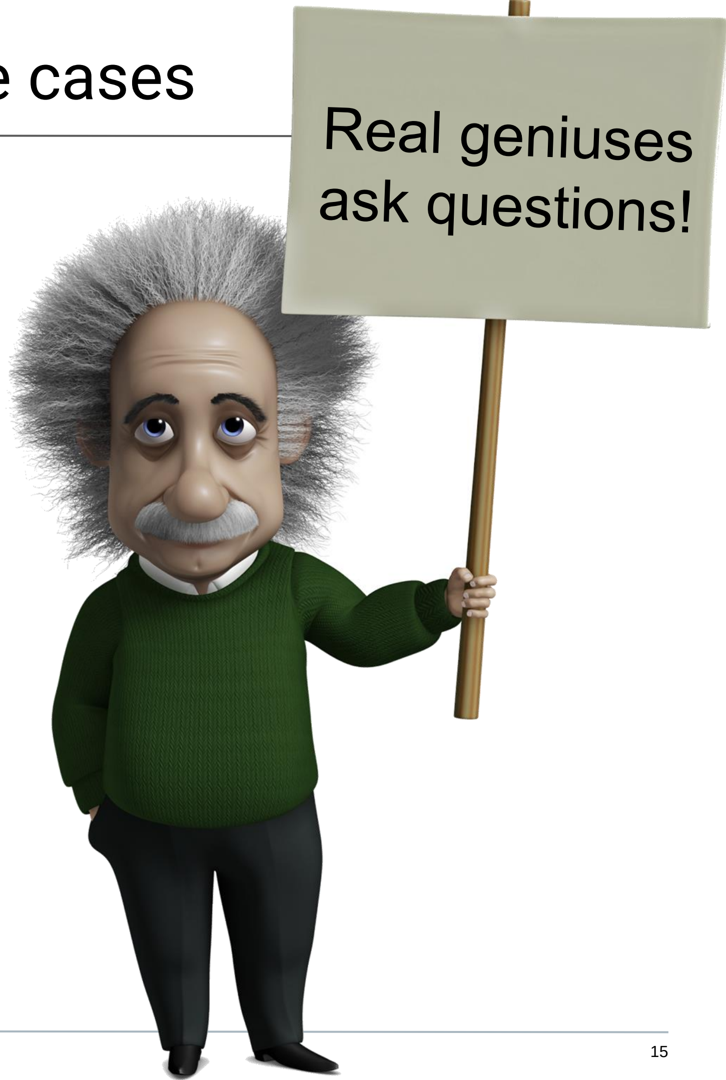




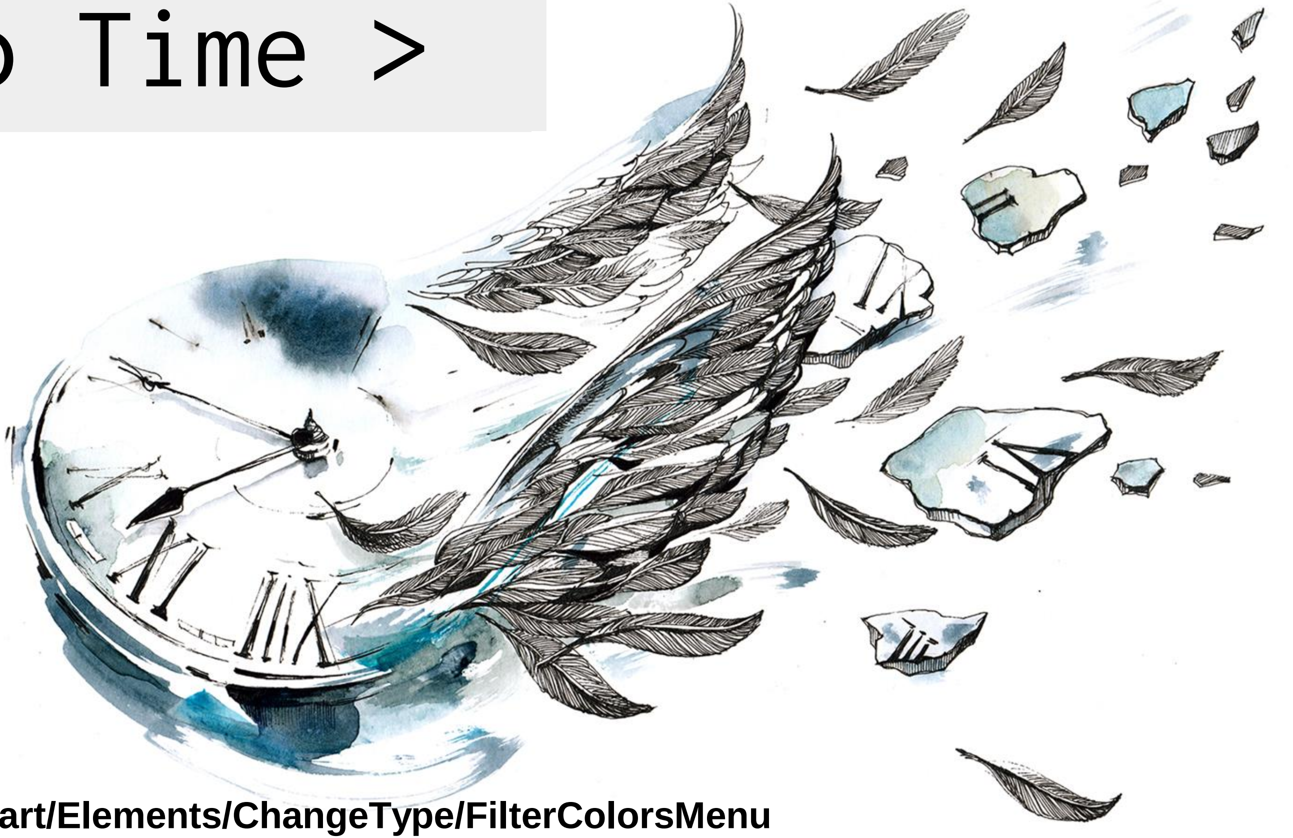


We will look at a few examples and use cases

- Try and follow along!
- In this activity we will
  - Look at 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 >



02-Ins IceCreamFaves(InsertBarChart/Elements/ChangeType/FilterColorsMenu /BasicCharts.2ndSheet-LineChart 1st

Activity: The Line and Bar Grades

03-Stu_LineAndBar/Student Grades – only 3 months of bar charts (any 3 indiv months)

Suggested Time:
15 Minutes

Activity: Line and Bar Grades

You are going to take the role of a teacher upon yourself for this activity as you create a series of bar and line graphs that visualize the grades of your class over the course of a semester.

Instructions:

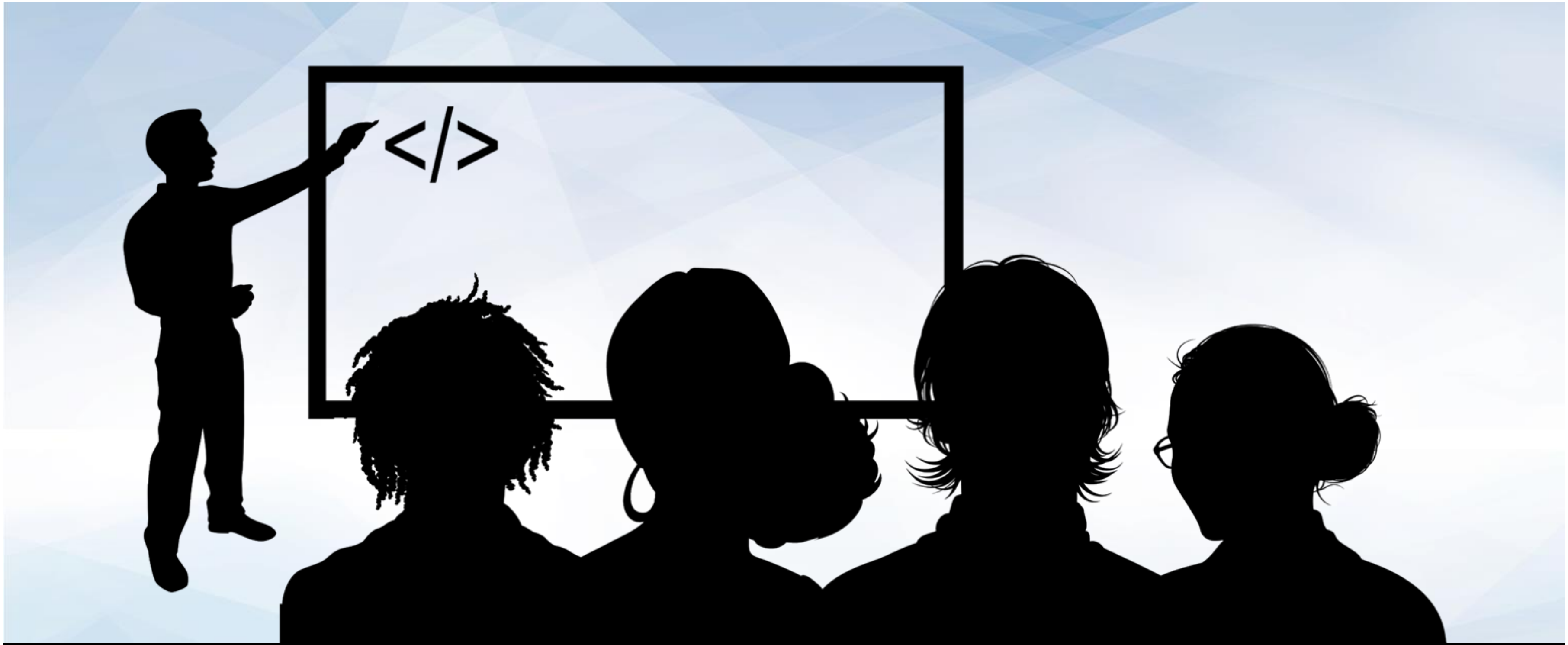
- Create a series of bar graphs that visualize the grades of all students in the class, one graph for every month.
- Create a line graph using all of the data that can be used to compare students' grades across the semester.
    - Use filtering in the line graph to allow you to drill down to a specific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

Hint:

- When duplicating bar graphs, it pays to get the formatting and look of the chart where you want it for the first graph (e.g. for January), and to then copy that chart and re-select the data for the subsequent copies (keeping the style and format, but just changing the data).

Suggested Time: 15 minutes

Time’s Up! Let’s Review.

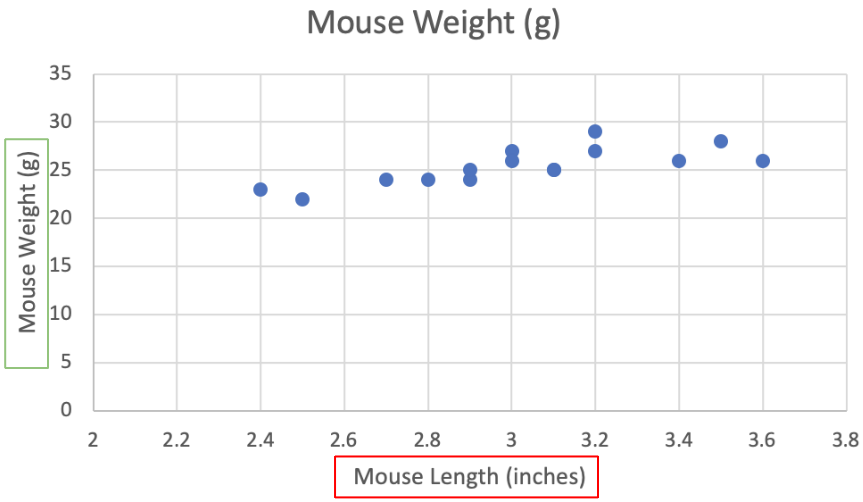


Instructor Demonstration

Scatter Plots and Trend Lines

Scatter plots are 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 >

04-Ins_ScatterPlot/ScatterPlot)

Partner Activity: Video Game Sales

05-Par_GameSales-ScatterPlots/VideoGameSales_Unsolved

Suggested Time:

15 Minutes

Partner Activity: Video Game Sales

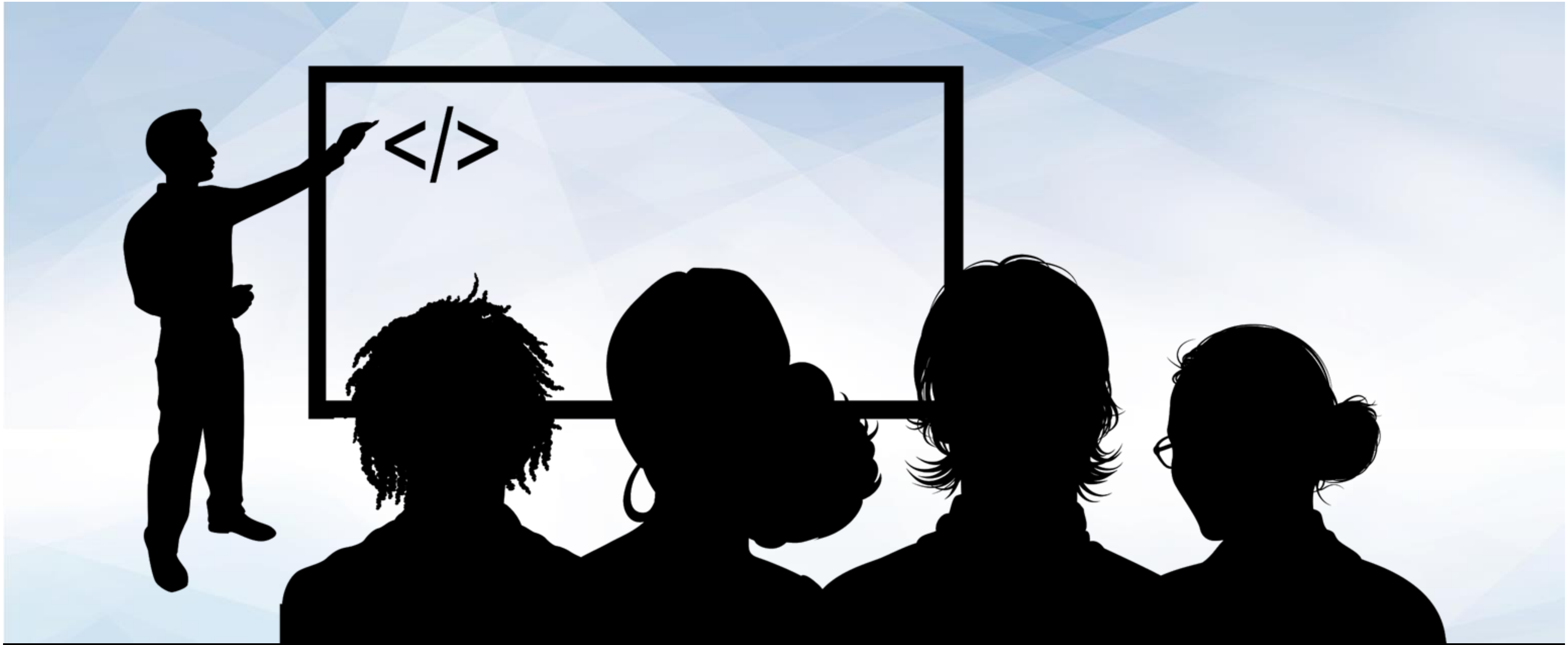
In this activity, you will pair up with one of your classmates in order to create a series of scatter plots which will compare video game sales across regions.

Instructions:

- Create a scatter plot that compares the NA (North American) sales of games versus the global sales of games. Make sure to add in axis titles, a chart title, and a trend line.
- Create a scatter plot that compares the EU (European) sales of games versus the global sales of games. Make sure to add in axis titles, a chart title, and a trend line
- Create a scatter plot that compares the JP (Japanese) sales of games versus the global sales of games. Make sure to add in axis titles, a chart title, and a trend line.
- Create a scatter plot that compares other sales of games versus the global sales of games. Make sure to add in axis titles, a chart title, and a trend line.
- Go back into each of your charts and modify the axes so that they are consistent for each chart.
  - Without consistency of margins between your charts they could be considered misleading.

Suggested Time: 15 minutes

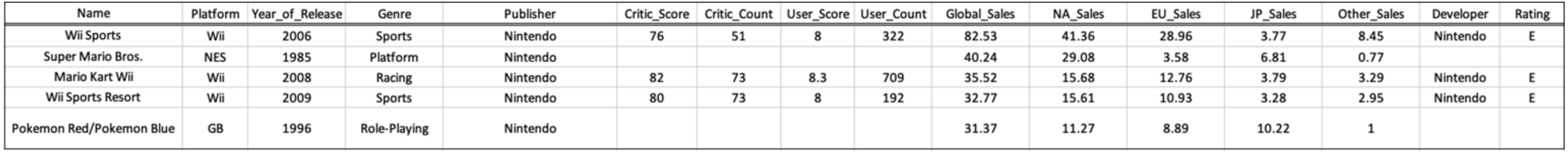
Time’s Up! Let’s Review.



Instructor Demonstration

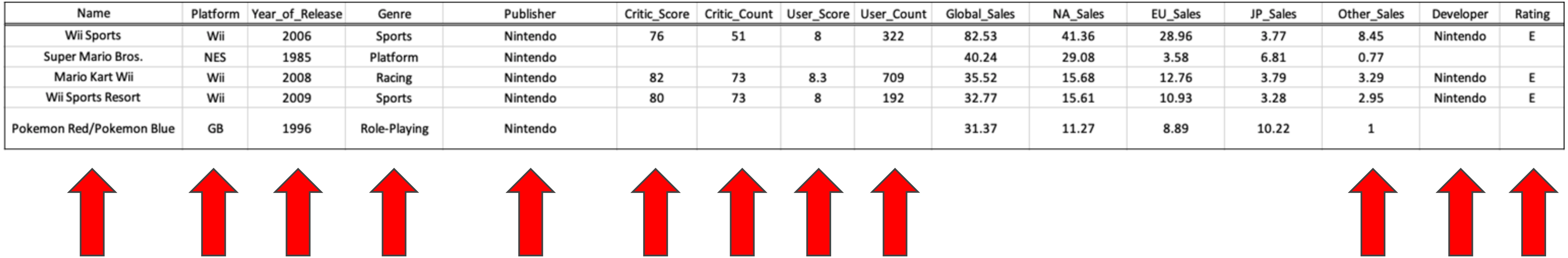
The Need to Filter

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



Name	Platform	Year_of_Release	Genre	Publisher	Critic_Score	Critic_Count	User_Score	User_Count	Global_Sales	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Developer	Rating
Wii Sports	Wii	2006	Sports	Nintendo	76	51	8	322	82.53	41.36	28.96	3.77	8.45	Nintendo	E
Super Mario Bros.	NES	1985	Platform	Nintendo					40.24	29.08	3.58	6.81	0.77		
Mario Kart Wii	Wii	2008	Racing	Nintendo	82	73	8.3	709	35.52	15.68	12.76	3.79	3.29	Nintendo	E
Wii Sports Resort	Wii	2009	Sports	Nintendo	80	73	8	192	32.77	15.61	10.93	3.28	2.95	Nintendo	E
Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo					31.37	11.27	8.89	10.22	1		

There was a **LOT** of unused data



Name	Platform	Year_of_Release	Genre	Publisher	Critic_Score	Critic_Count	User_Score	User_Count	Global_Sales	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Developer	Rating
Wii Sports	Wii	2006	Sports	Nintendo	76	51	8	322	82.53	41.36	28.96	3.77	8.45	Nintendo	E
Super Mario Bros.	NES	1985	Platform	Nintendo					40.24	29.08	3.58	6.81	0.77		
Mario Kart Wii	Wii	2008	Racing	Nintendo	82	73	8.3	709	35.52	15.68	12.76	3.79	3.29	Nintendo	E
Wii Sports Resort	Wii	2009	Sports	Nintendo	80	73	8	192	32.77	15.61	10.93	3.28	2.95	Nintendo	E
Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo					31.37	11.27	8.89	10.22	1		

- 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

06-Evr_PigeonRacing-Filter/Unsolved/PigeonRacing

Partner Activity: Video Game Sales

07-Par_FilterGameSales/VideoGameSales2_Solved

Suggested Time:
15 Minutes

Partner Activity: Video Game Sales

Instructions:

- Create a scatter plot which graphs the critical response (Critic Score) of games published by Nintendo as compared to their global sales.
- Create a scatter plot which graphs the critical response of games published by Electronic Arts as compared to their global sales.
  - Only chart those games that have been reviewed. Games without any reviews should be ignored.
  - Add a chart title, axis titles, and a trend line to the graph that is created.
- Select all of the data on the worksheet and create a line chart which can be filtered by publisher, whose rows are set by a game's year of release, and whose values are the sum of global sales for that year.
  - Create a 2D line graph that charts this data.

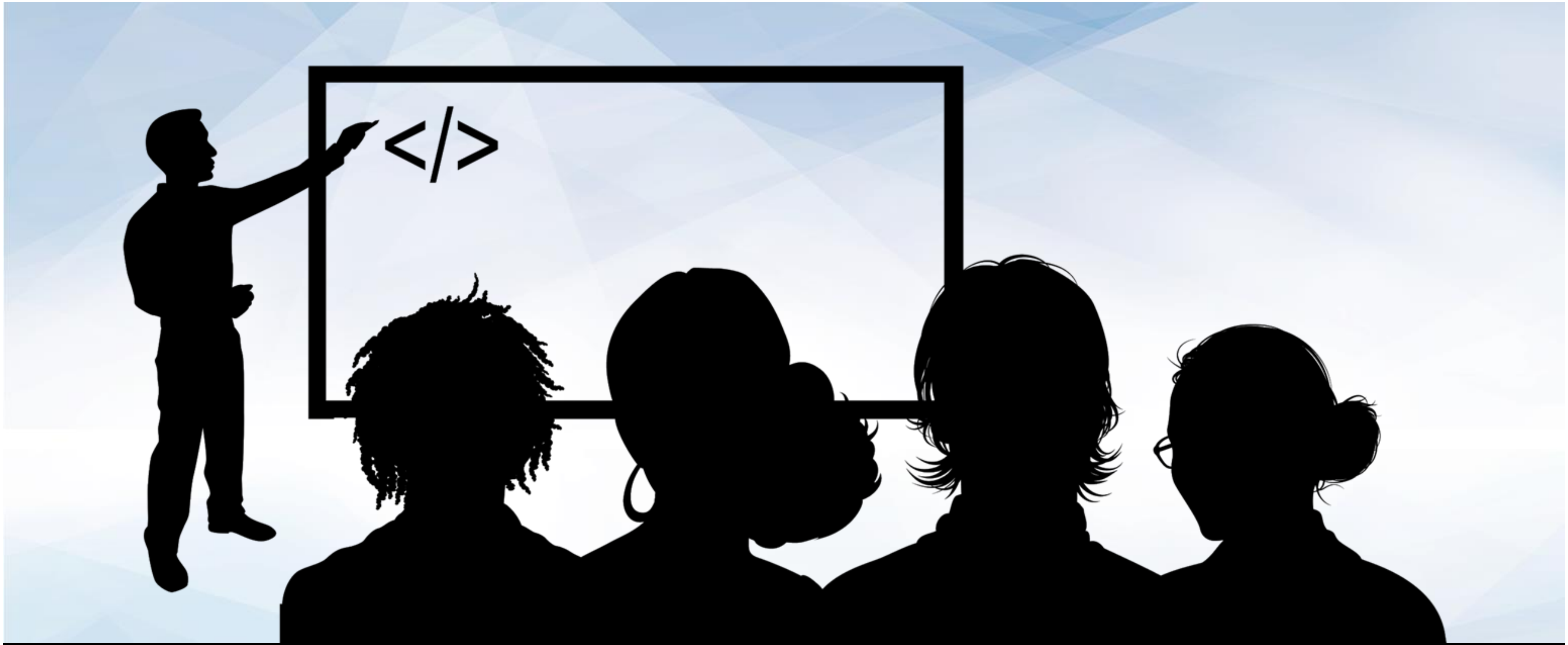
Notes:

- Only chart those games that have been reviewed. Games without any reviews should be ignored.
- Add a chart title, axis titles, and a trend line to the graph that is created.

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.

What are the measures of central tendency used for?

A

Metrics used to describe the center of a data set.

How do you describe the variability of a data set?

Three summary statistics metrics for describing variability

01 Variance

02 Standard Deviation

03 Z-Score

Variance

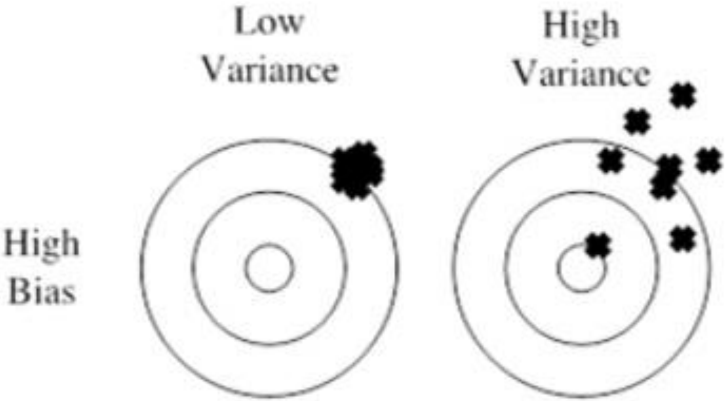
- Used to describe how far values in the data set are from the mean
- Describes how much variation exists in the data
- Variance considers the distance of each value in the data set from the center of the data

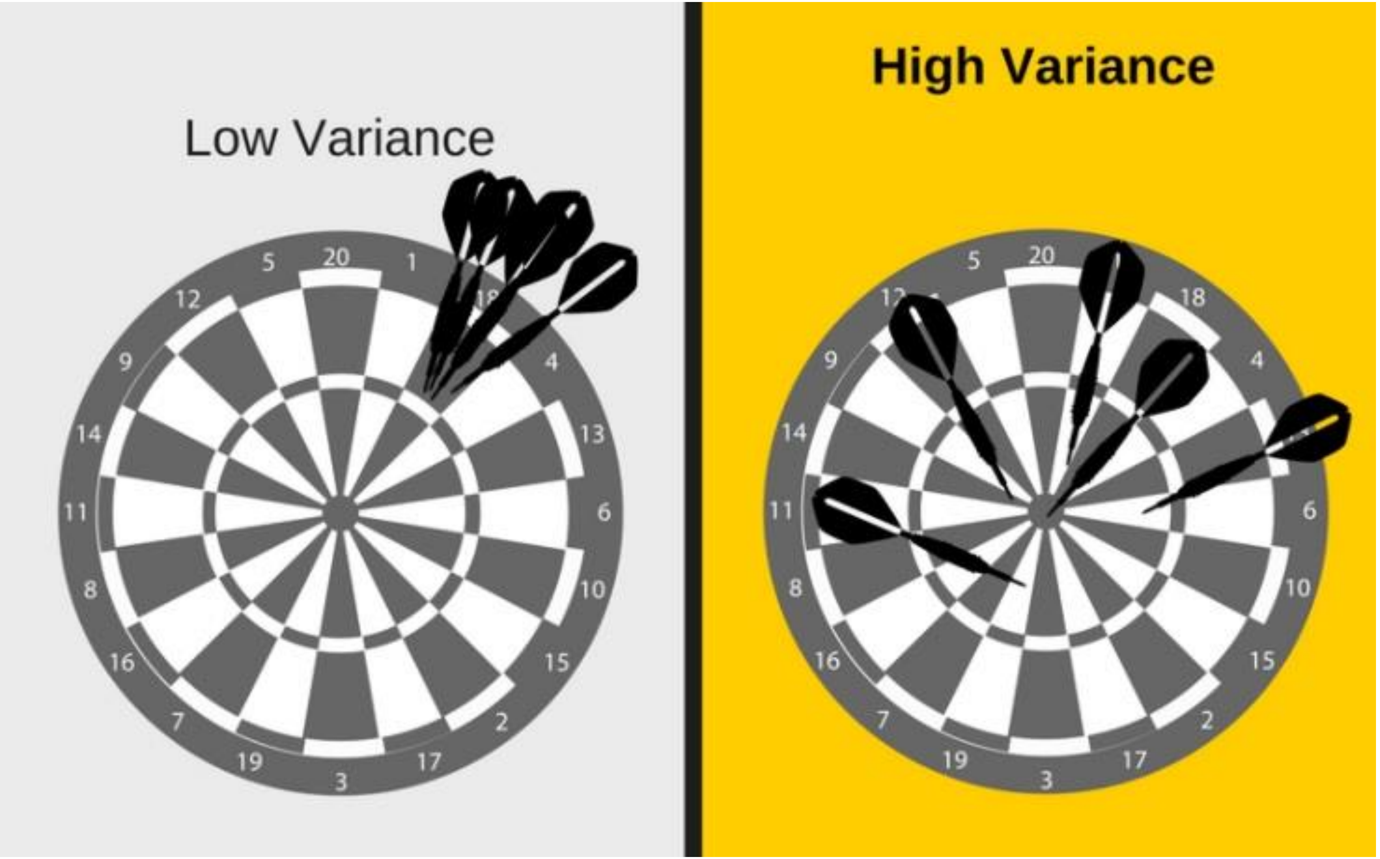
Variance = Sum[(Datapoint-MeanOfDataSet)^2]
Number of Datapoints in Sx

- 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

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

Dart Example




Low Variance
20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5
High Variance
20
18
4
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5

<Time to calculate variance>

Standard Deviation

- Describes how *spread out* the data is from the mean
- Calculated from the square root of the variance
- In the same units of measurement as the mean

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

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

<Time to calculate standard deviation>

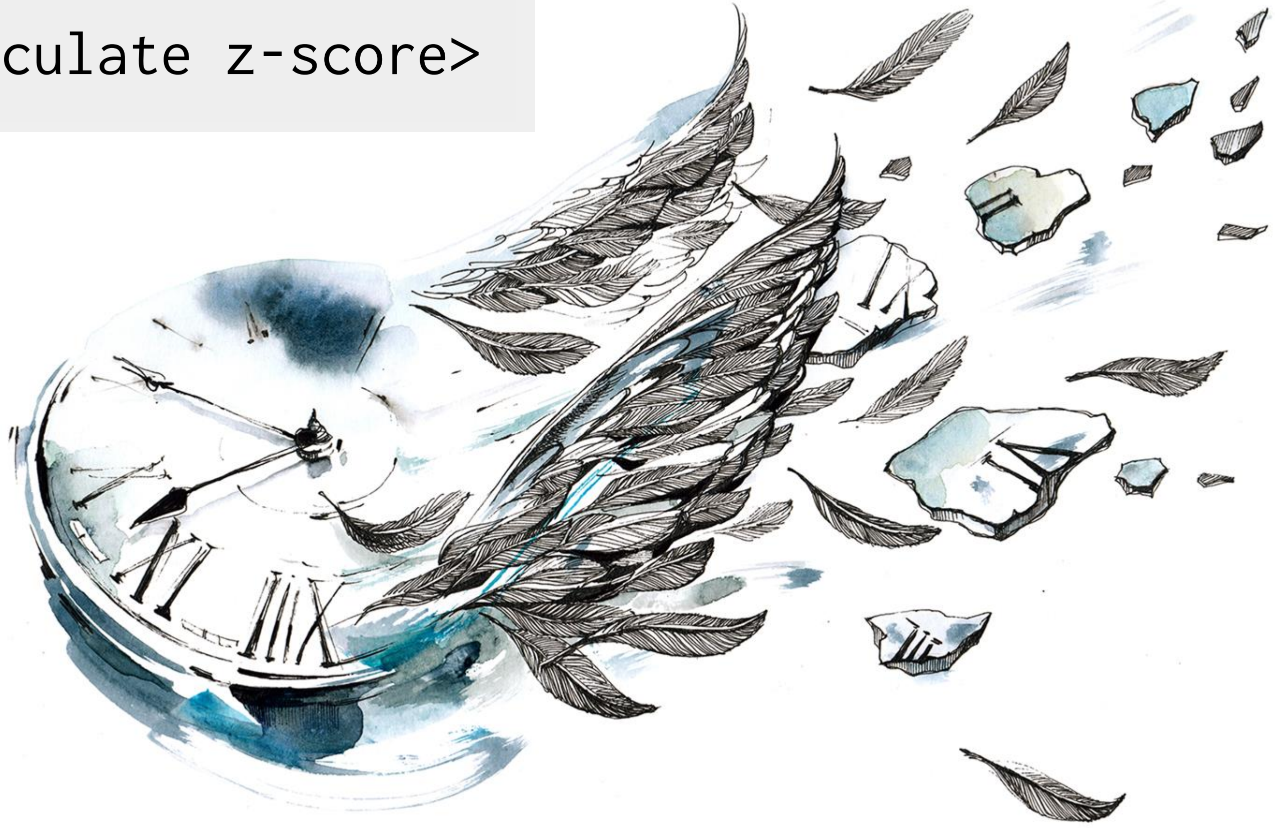
Z-Score

- Describes a single value’s distance from the mean of the data set
- The distance is in terms of standard deviations
- 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

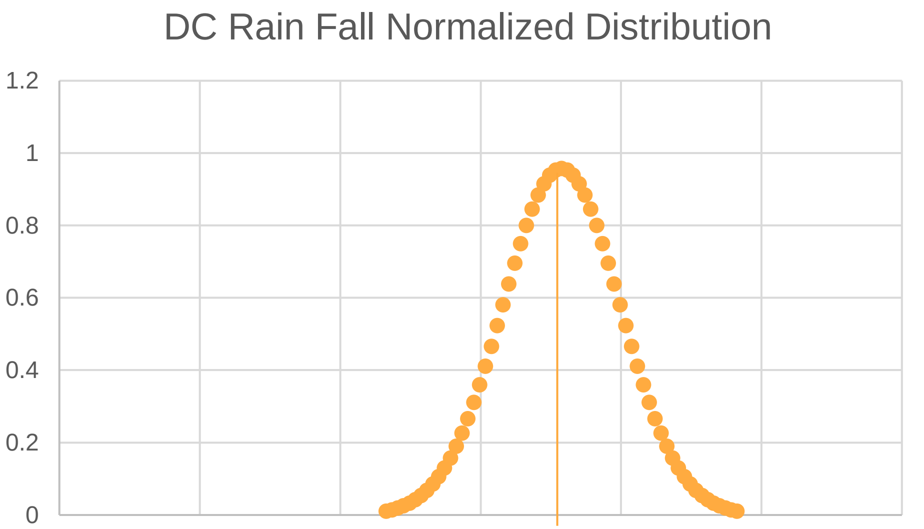
$$z = \frac{X - \mu}{\sigma}$$

<Time to calculate z-score>

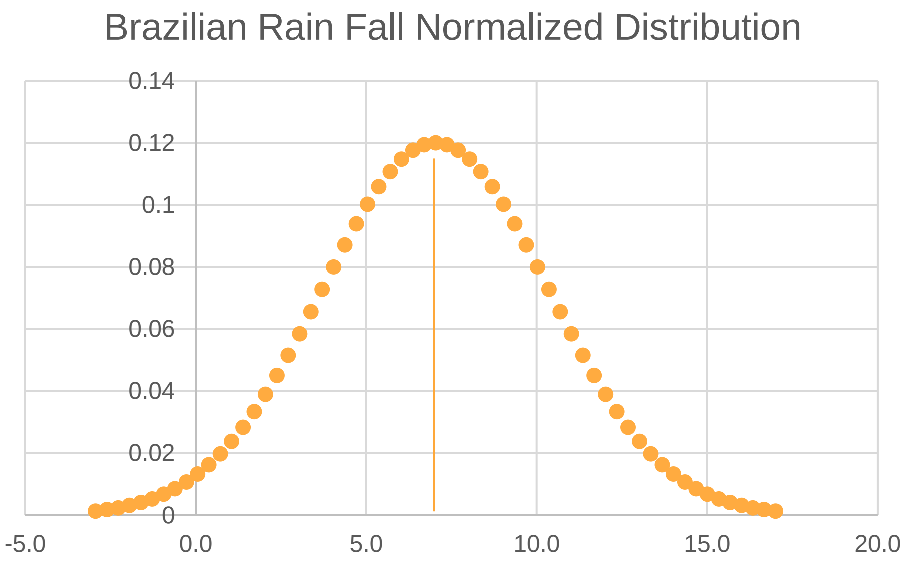


Activity: Variance, Standard Deviation and Z-Score Review

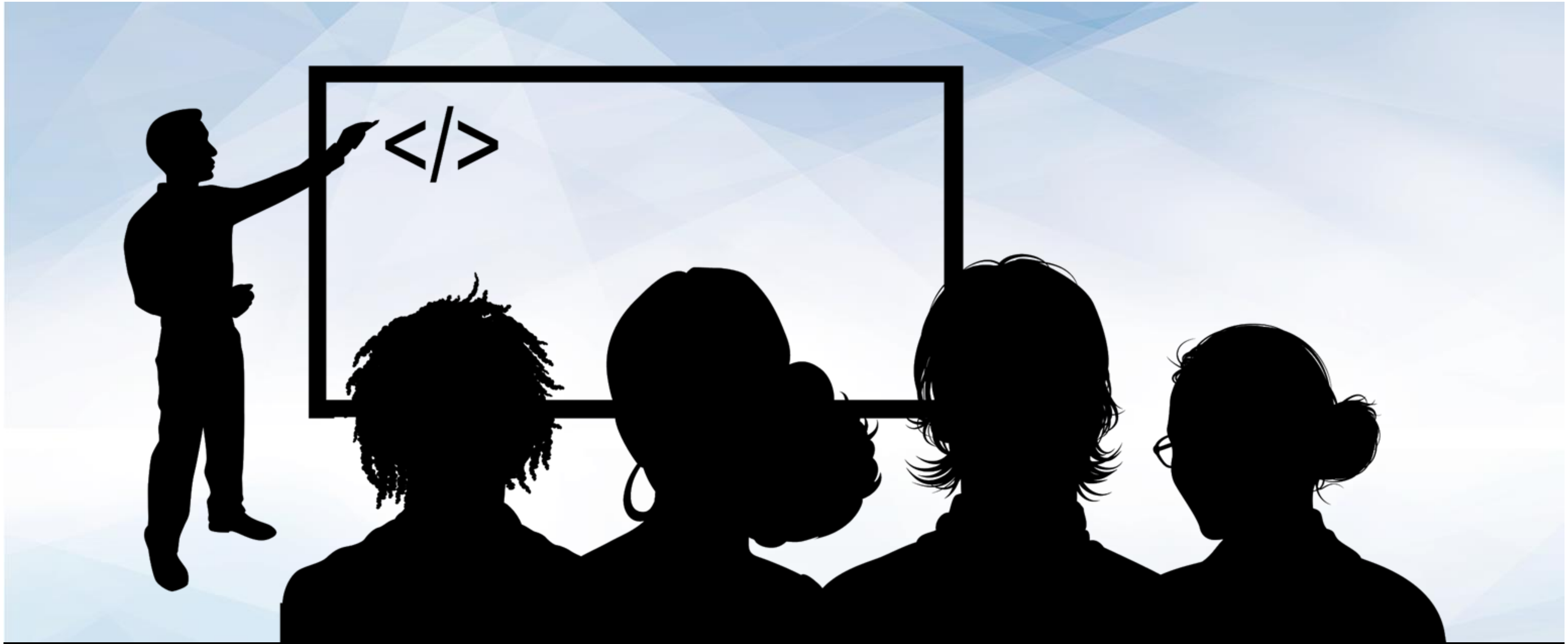
Suggested Time:
15 Minutes


DC Rain Fall Normalized Distribution
1.2
1
0.8
0.6
0.4
0.2
0
0.0
1.0
2.0
3.0
4.0
5.0
6.0

Mean 3.6


Brazilian Rain Fall Normalized Distribution
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-5.0
0.0
5.0
10.0
15.0
20.0

Mean 7.0



Instructor Demonstration

Quantiles, Outliers and Boxplots

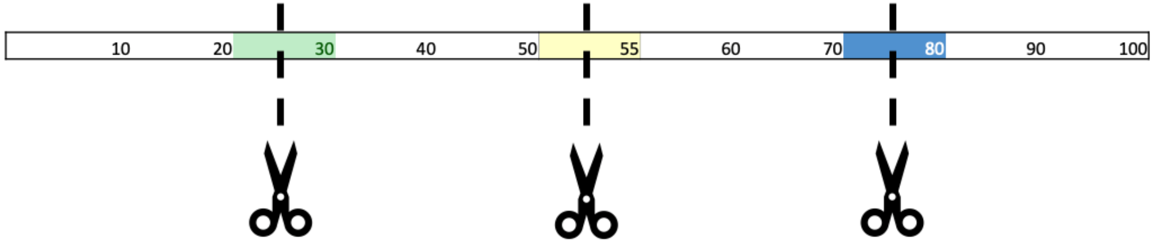
Be careful when describing real-world data

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

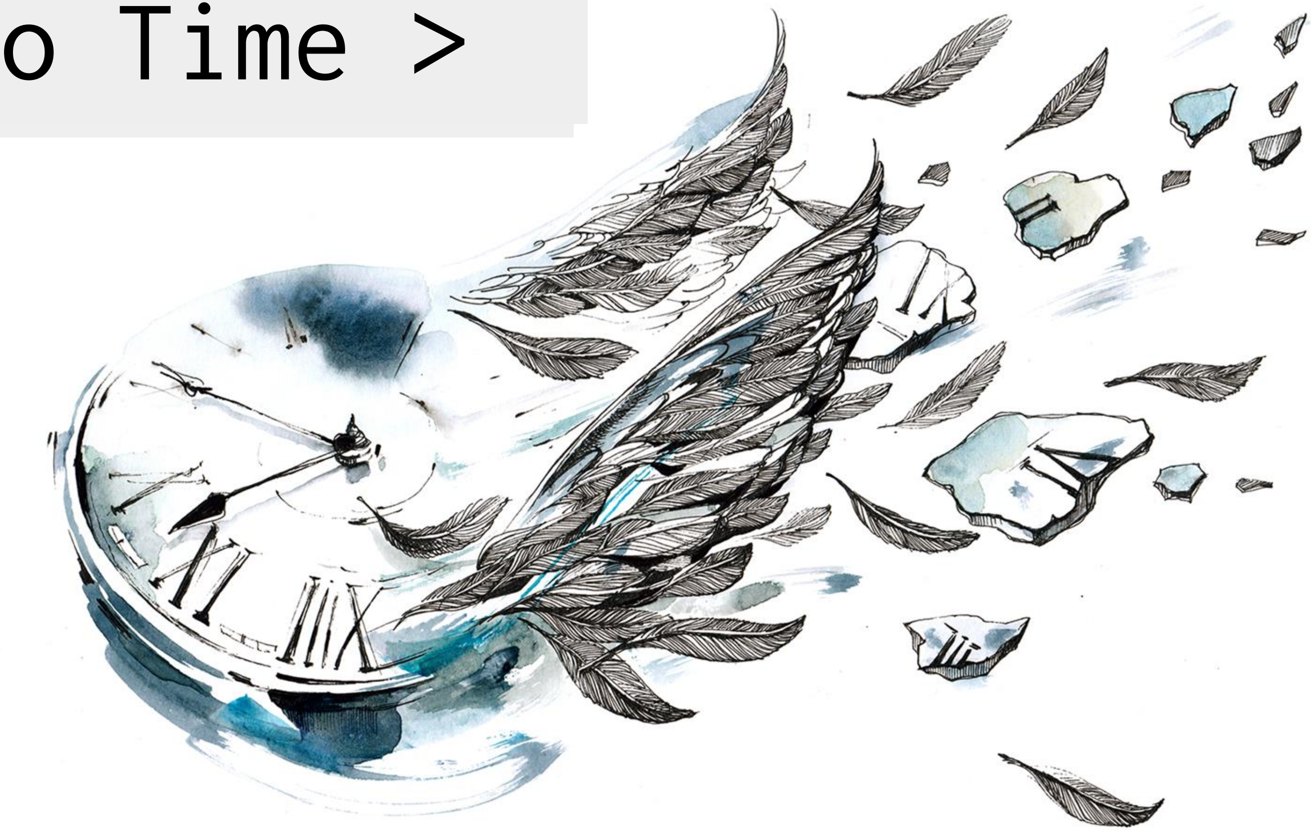
But how can we summarize real- world data?

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 >



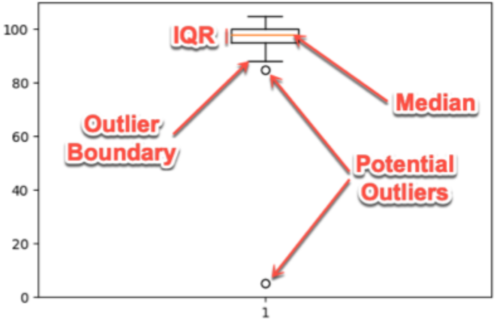
Extreme values may not always be reliable

- **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
  - Can cause us to misrepresent the actual data

There are two ways to identify potential outliers

01 Qualitatively

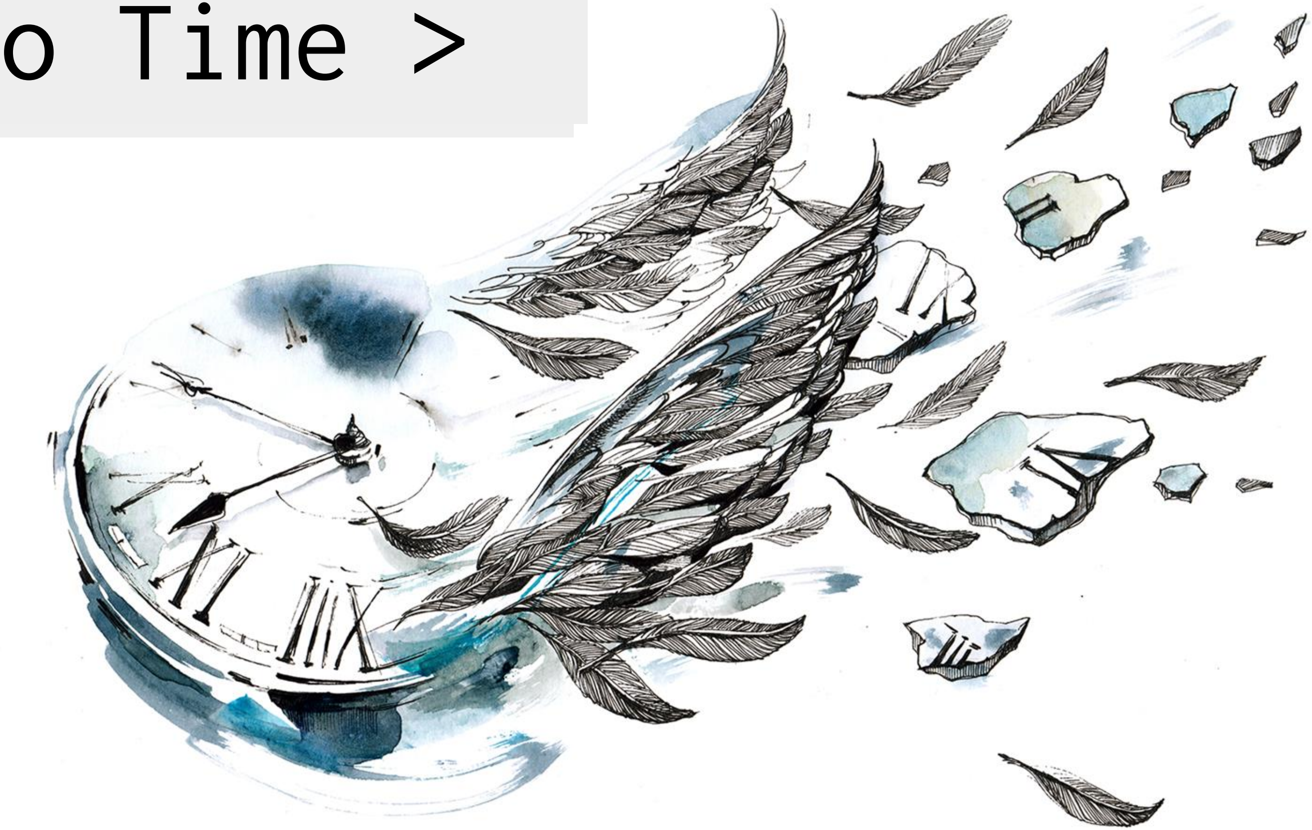
- Use box and whisker plots to visually identify potential outlier data points



02 Quantitatively

- Determine the outlier boundaries in a dataset 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

Suggested Time:
10 Minutes

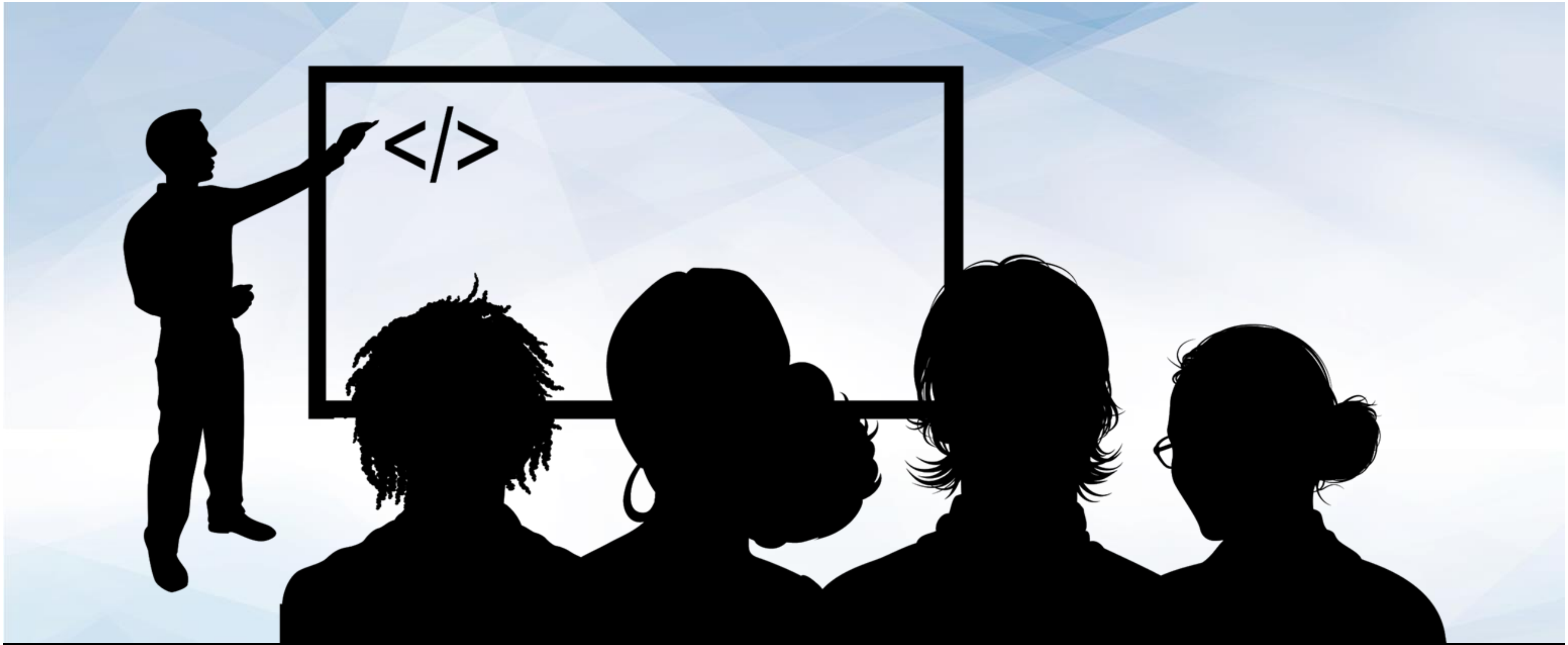
Variance, Standard Deviation and Z-Score Review Instructions

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

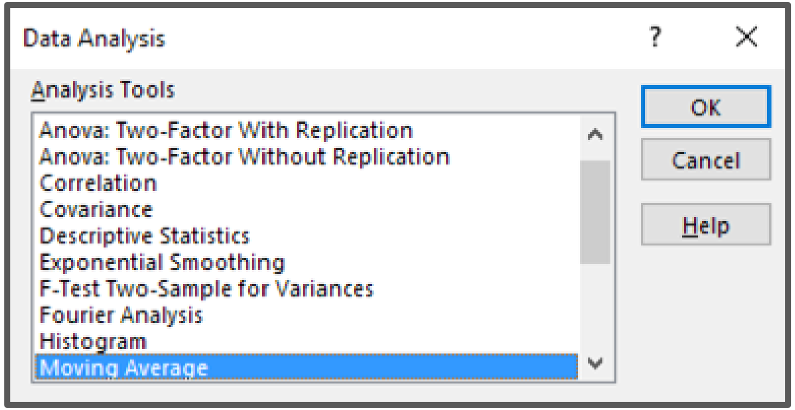
Excel’s Statistics Add-On

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
- All of these functions we will cover throughout the course!



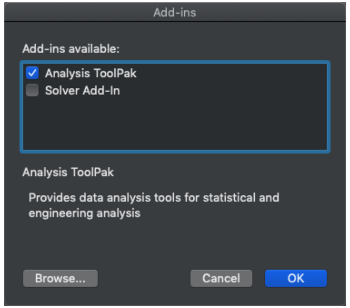
Analysis ToolPak is not designed for in-depth data analytics

- 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
  - Gut-checks
  - One-off analysis

How to install and use the Excel Analysis ToolPak - Mac

To Install:

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



Add-ins available:

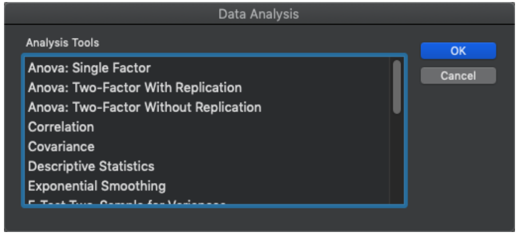
☑ Analysis ToolPak
□ Solver Add-In

Analysis ToolPak

Provides data analysis tools for statistical and engineering analysis

To Use:

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



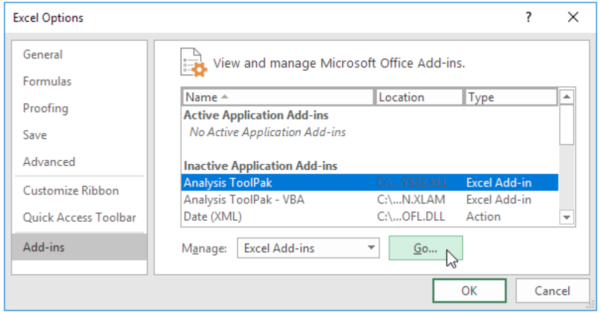
Analysis Tools

Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances

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.



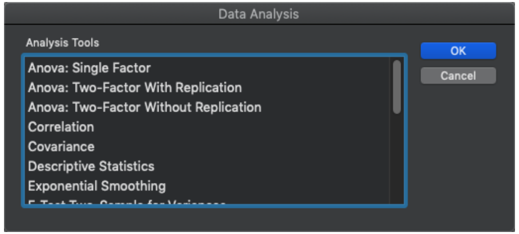
View and manage Microsoft Office Add-ins.

Name	Location	Type
Active Application Add-ins		
No Active Application Add-ins		
Inactive Application Add-ins		
Analysis ToolPak	C:\...YS32.XLL	Excel Add-in
Analysis ToolPak - VBA	C:\...N.XLAM	Excel Add-in
Date (XML)	C:\...OFL.DLL	Action

Manage: Excel Add-ins

To Use:

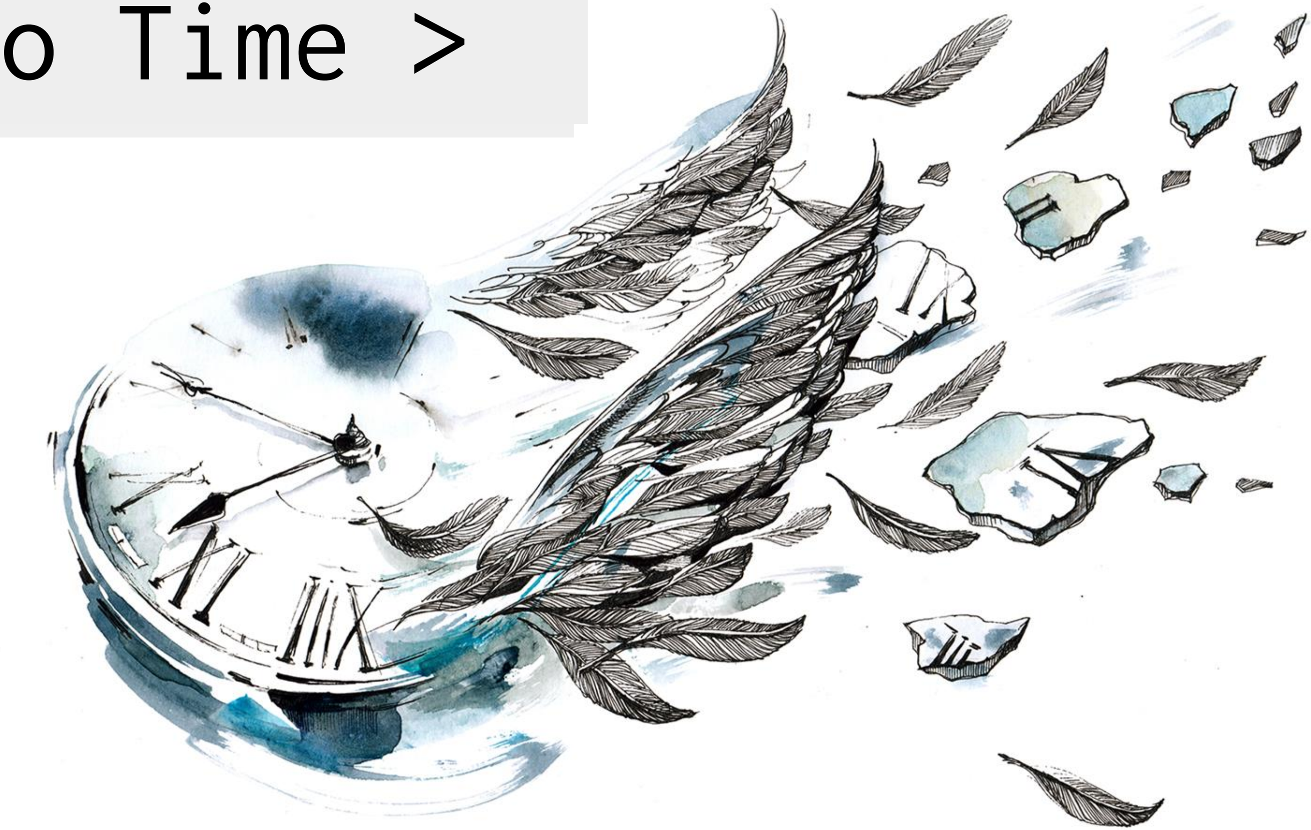
1. Go to the “Data” menu in Excel.
2. Go to the “Analyze” section.
3. Select the “Data Analysis” option.

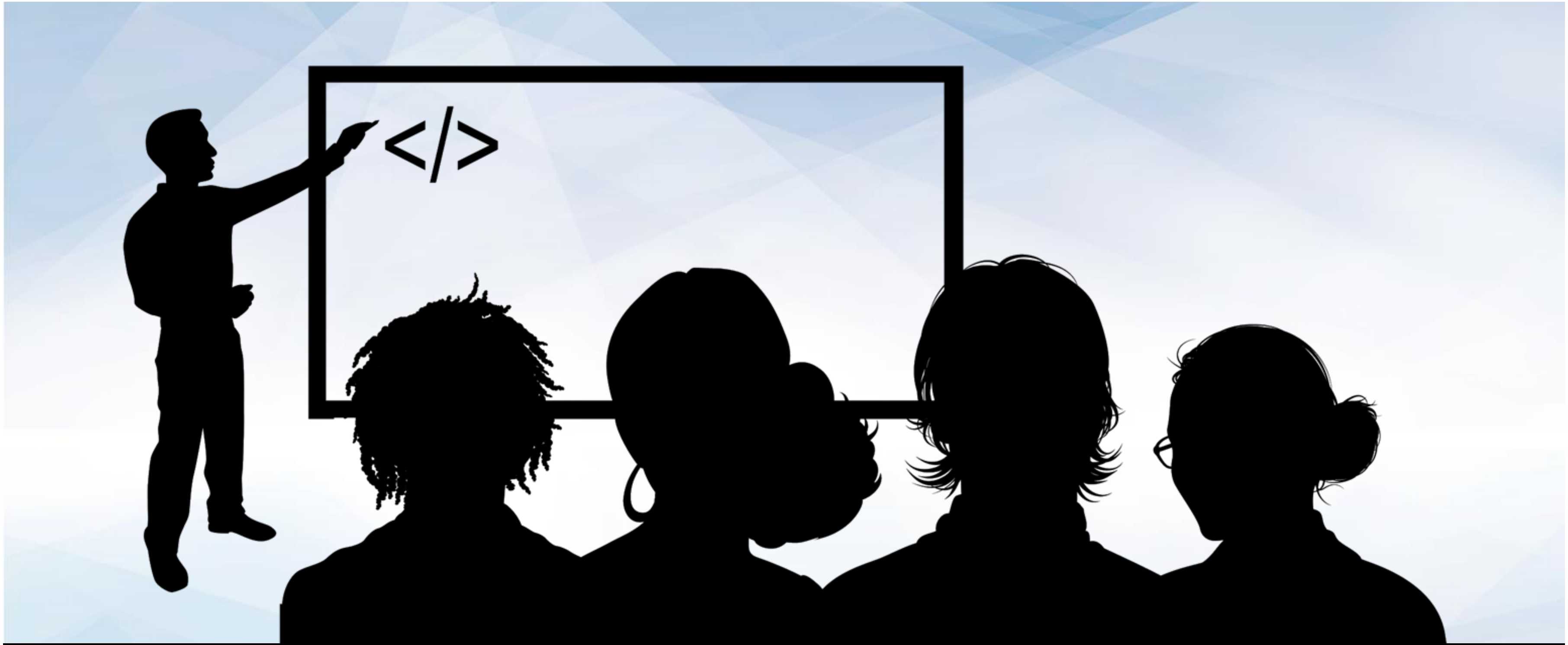


Analysis Tools

Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances

< Demo Time >





Instructor Demonstration

Adding Files to Github

Github is a hosting service for source code

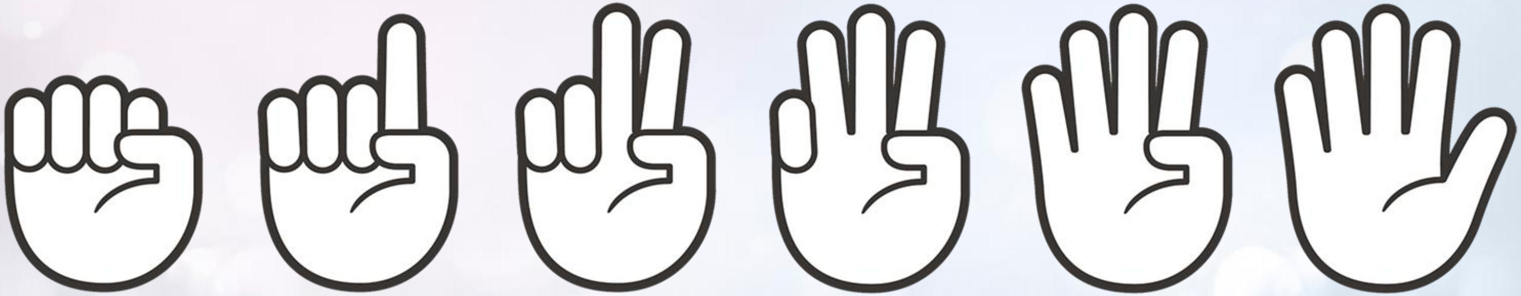
- Web interface for **Git**
- **Git** is version control software
  - Tracks source code history
  - Allows for collaboration on the same code files across a team or organization
  - Easily update and rollback software versions
- Since 2019, Github is used by over 2.1 million companies
- Proficiency in Git and Github is 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 functionality.

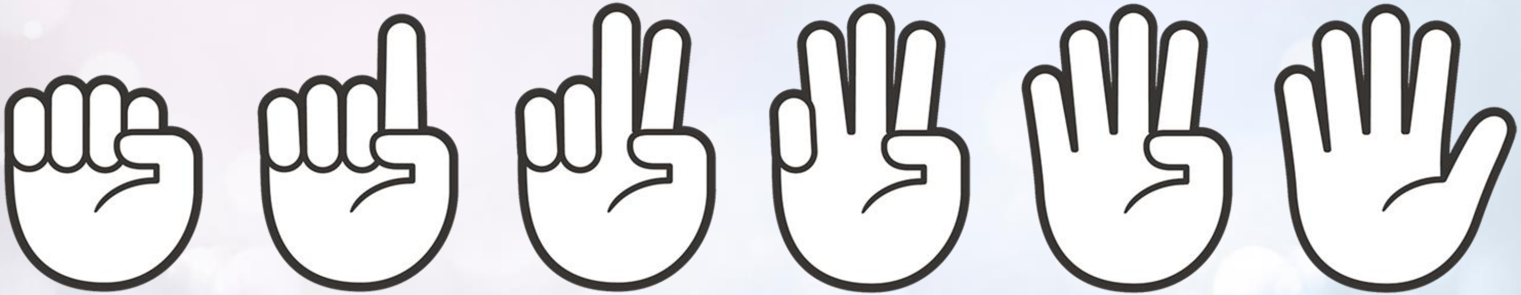
< Demo Time >

Github/GitLab



FIST TO FIVE:

Who feels comfortable with plotting figures in Excel?



FIST TO FIVE:

Who feels comfortable calculating summary statistics in Excel?