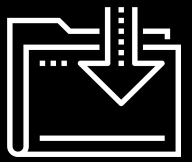
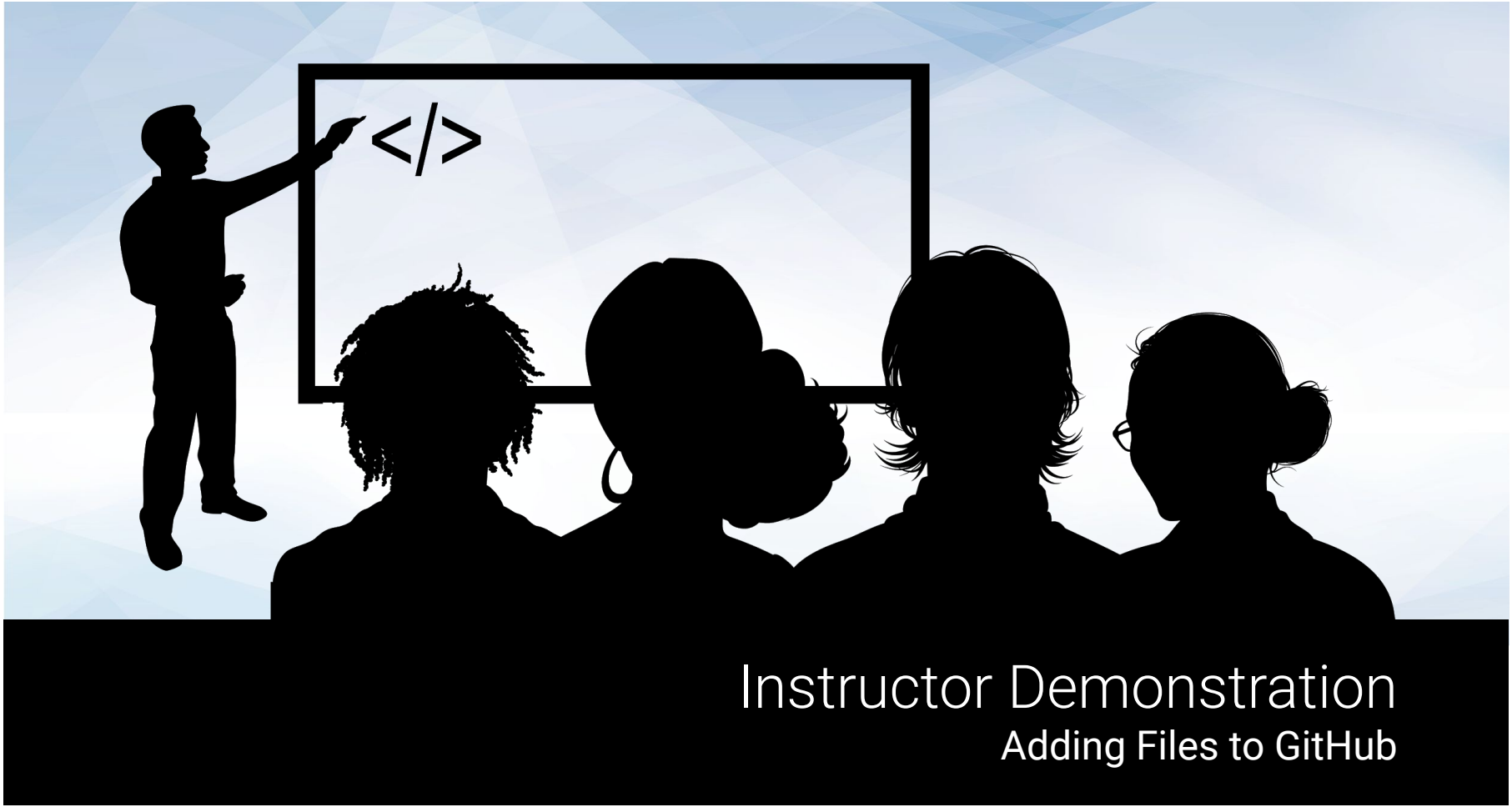
Excel Plotting

Data Boot Camp

Lesson 1.3




WELCOME



Instructor Demonstration

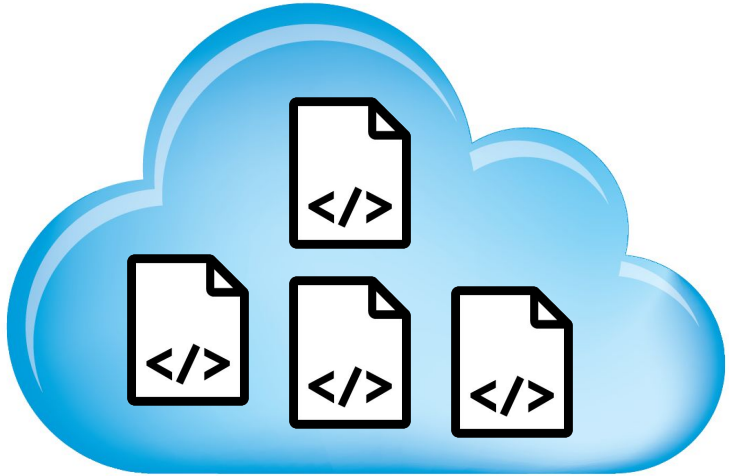
Adding Files to GitHub

GitHub Is a Hosting Service for Source Code

GitHub is a web interface for Git

Git is version control software that can:

- Track source code history
- Allow for collaboration on the same code files across a team or organisation
- Easily update and rollback software versions



Since 2019, GitHub is 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

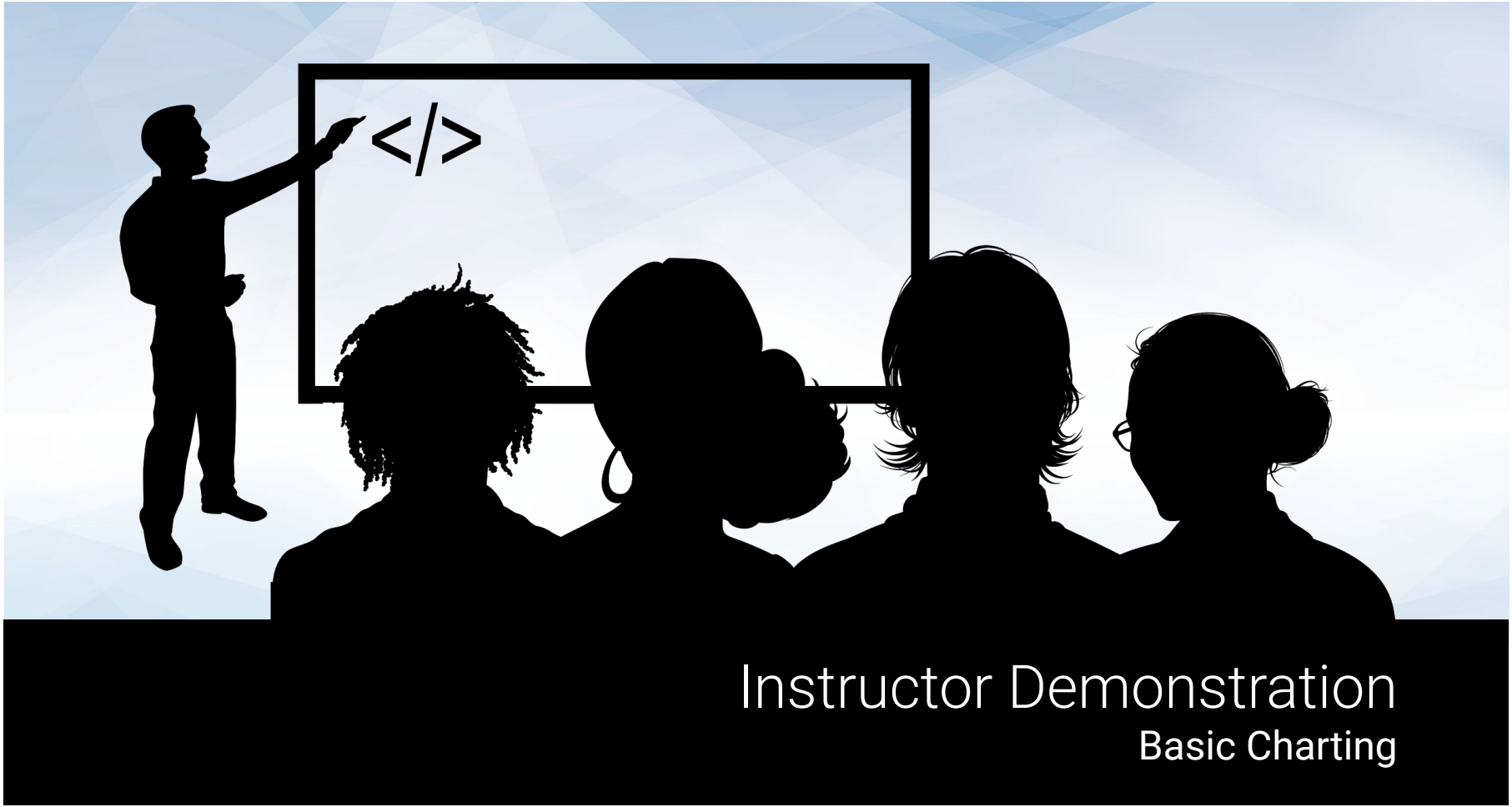
You will submit your homework assignments using GitHub

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 functionality

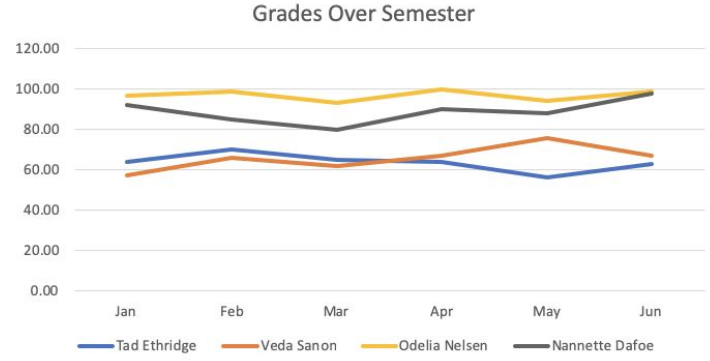
< Demo Time >

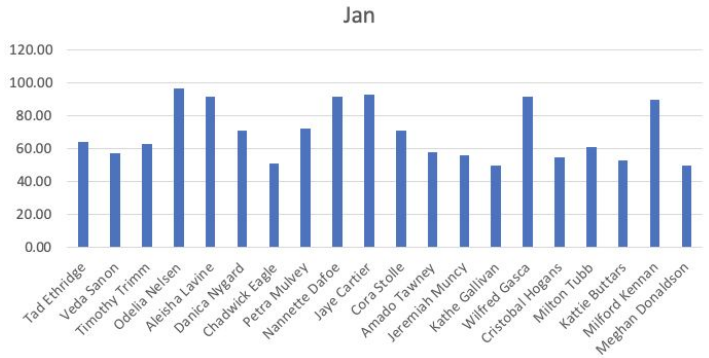


Instructor Demonstration

Basic Charting

It Is Time to Learn Excel Visualisations!





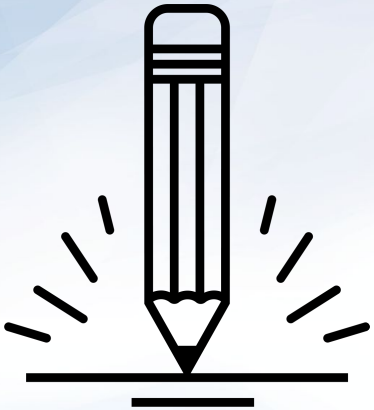
We Will Look at a Few Examples and Use Cases

In this activity we will:

- Look at an example data set
- Select data of interest
- Visualise selected data
- Add labels and titles to our visualisation

Do not hesitate to ask questions.
Our TAs will slack out images for each operating system

< Demo Time >



Activity: The Line and Bar Grades

Suggested Time:
15 Minutes

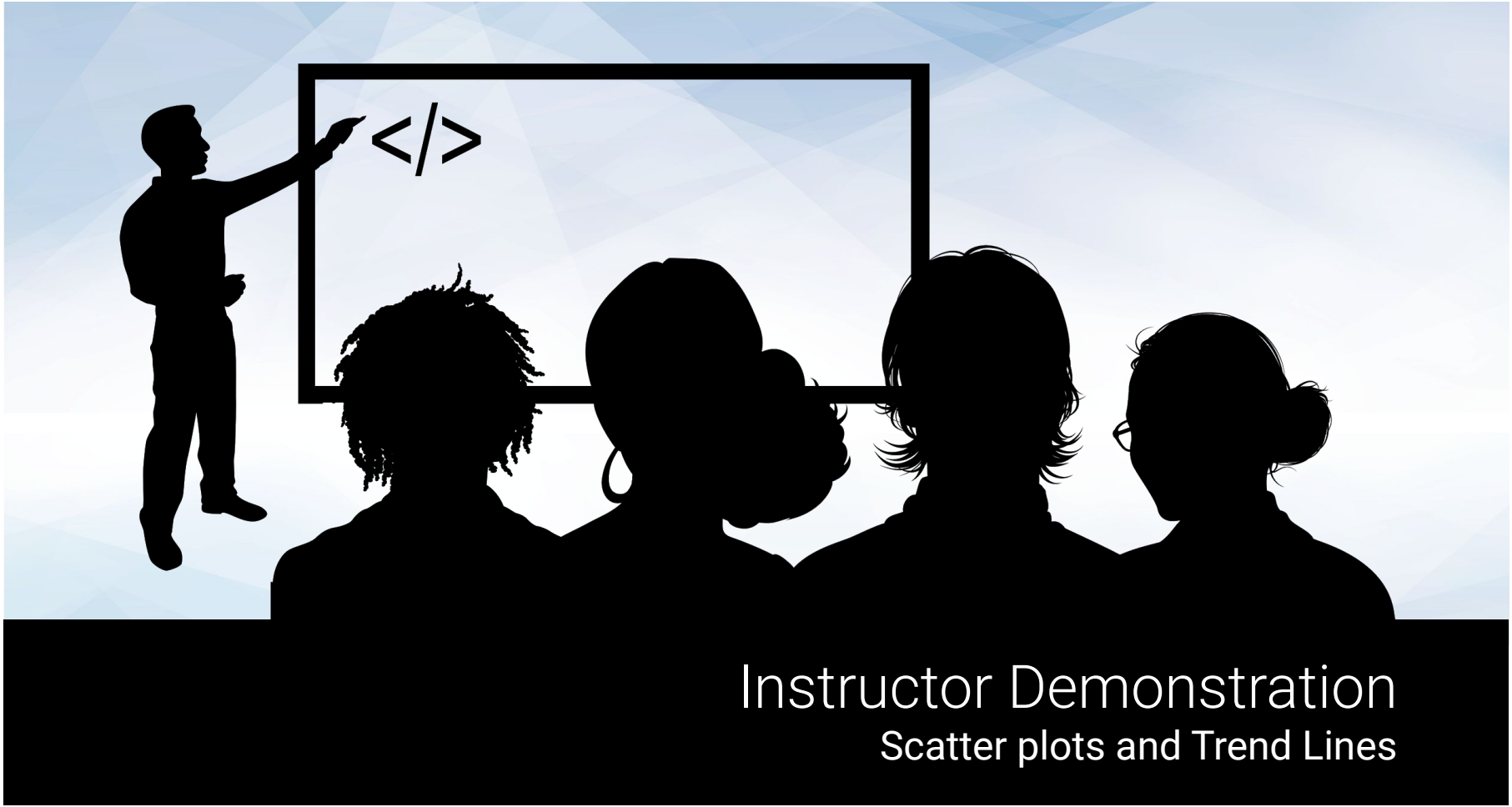
Activity: Line and Bar Grades

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 visualise the grades of your class over the course of a semester.

Instructions:	**Hint:**
• Create a series of bar graphs that visualise the grades of all students in the class, with 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.	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.



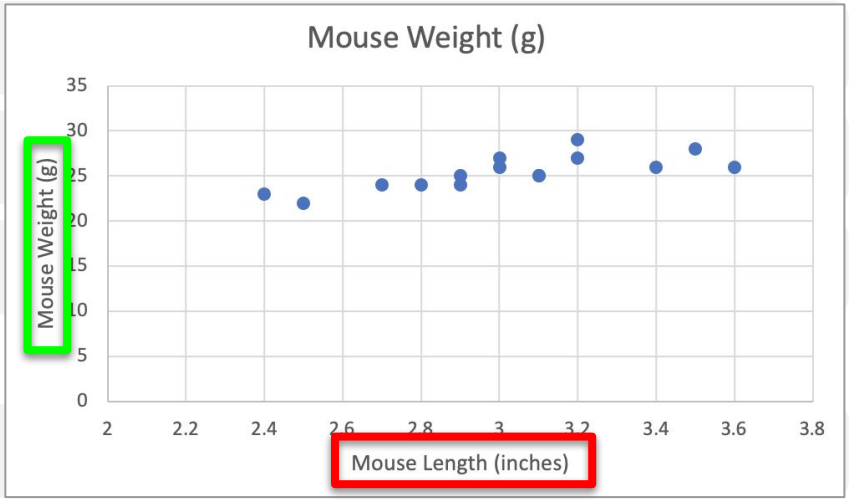
Instructor Demonstration

Scatter plots and Trend Lines

Scatter Plots Are a Powerful Visualisation Tool!

Visualises 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 scatter plot can make visualising this relationship even easier!

< Demo Time >

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 different regions.

Suggested Time:
15 minutes

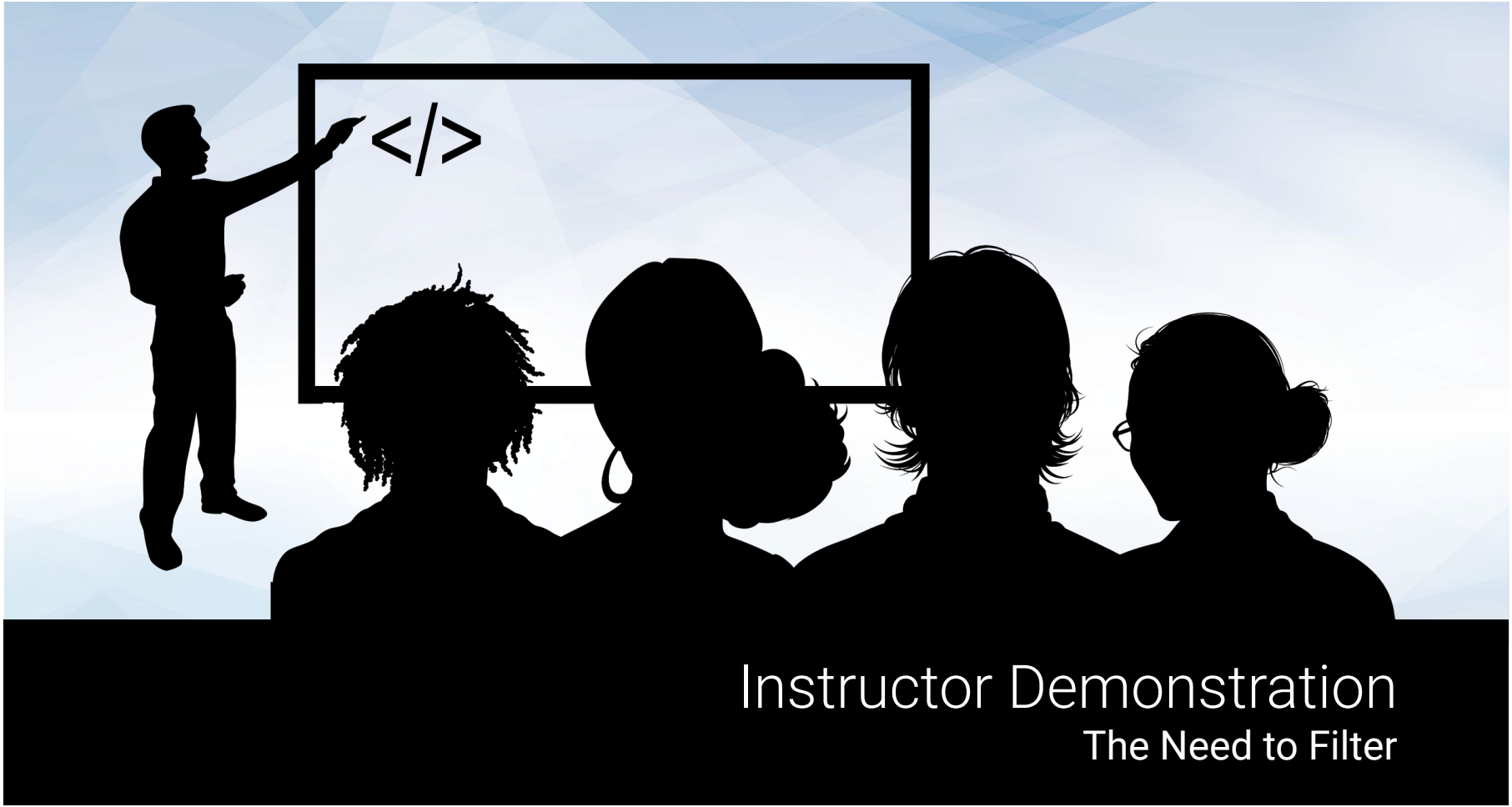
Partner Activity: Video Game Sales

Instructions:

- Create a scatter plot that compares AUS (Australian) 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 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.
- 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.

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

There Was a LOT of Unused Data

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

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: Filter Game Sales

Suggested Time:
15 minutes

Partner Activity: Filter 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.

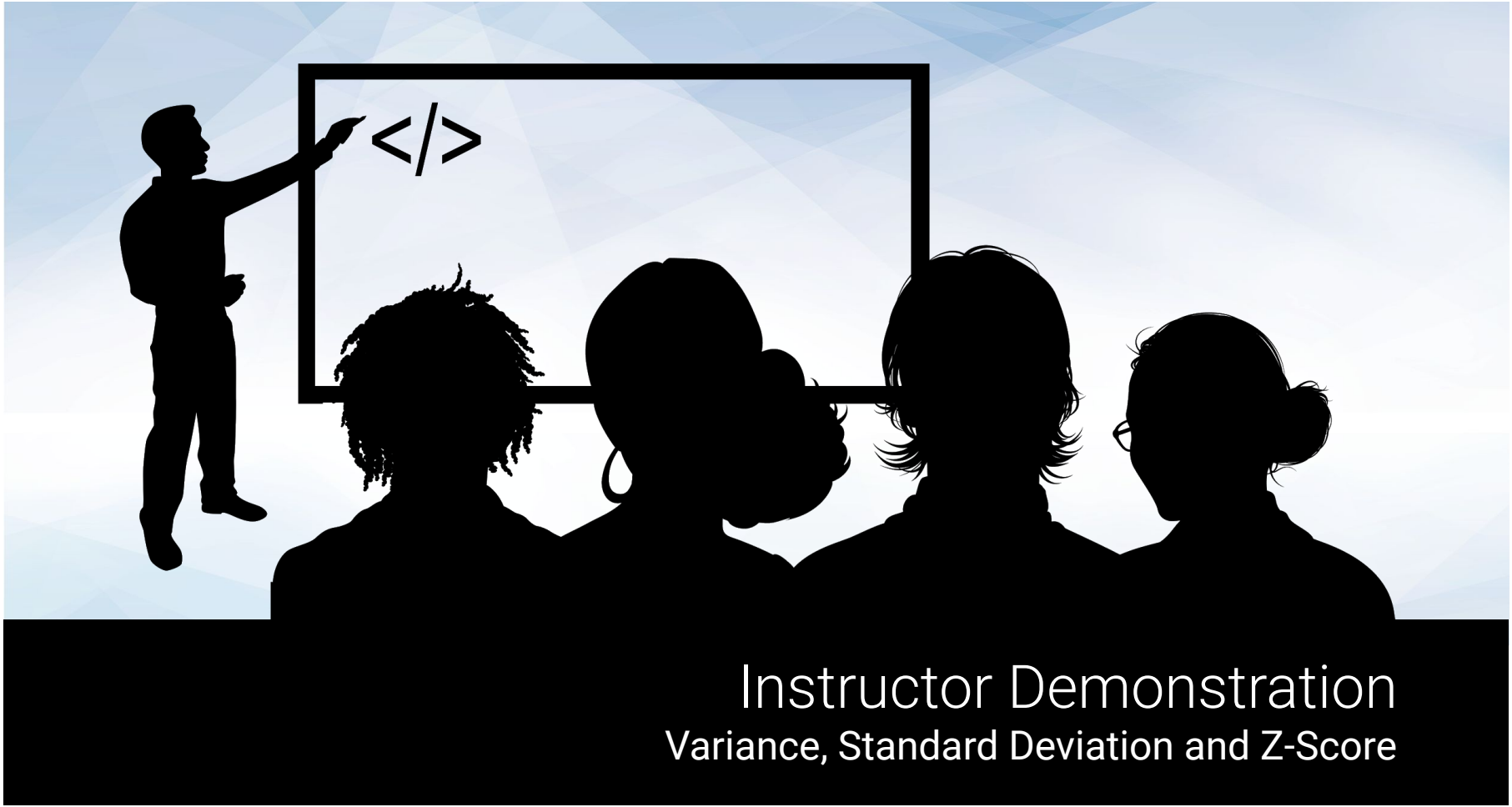
Notes:

- 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.


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?

Metrics used to describe the centre 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
- Considers the distance of each value in the data set from the centre of the data

σ2	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

- 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
σ2	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

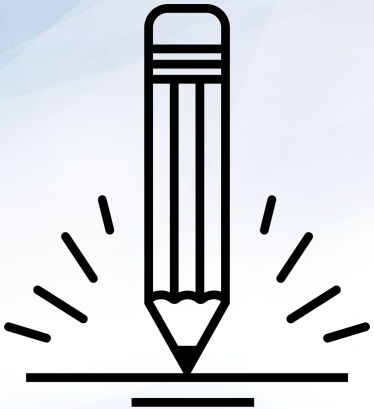
Can be positive or negative
- If negative, the value is less than the mean
- If positive, the value is greater than the mean

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>



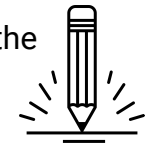
Activity: Variance, Standard Deviation, and Z-Score Review

Suggested Time:
15 Minutes

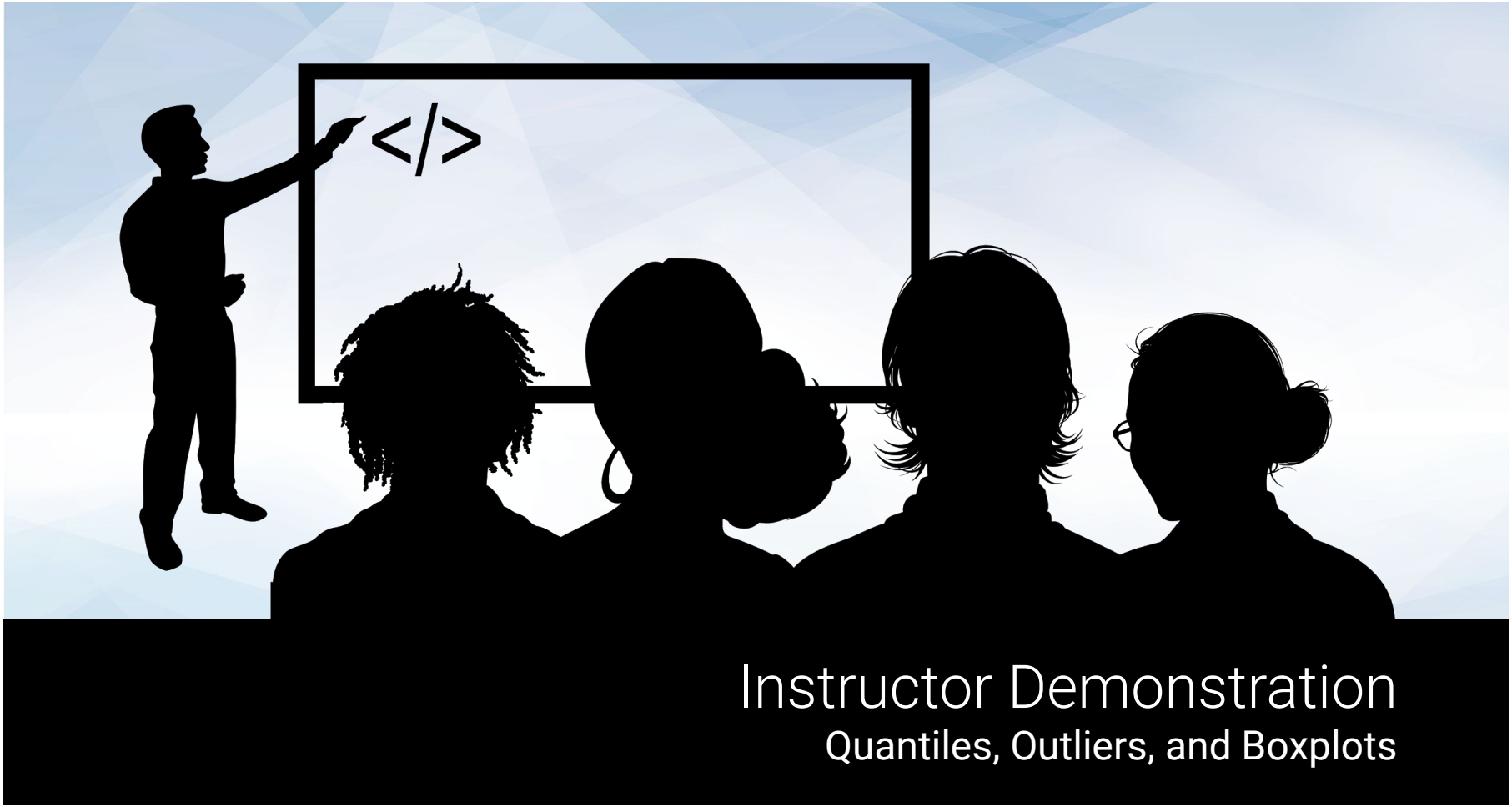
Variance, Standard Deviation, and Z-Score Review Instructions

- Open the workbook that contains your raw data.
  - File: `Unsolved/variance_review.xlsx`
- Create a new sheet in the workbook and name the sheet ‘Summary Table’.
- Within the new sheet, create a Team column, which contains the following teams:
  - CLE, GSW, LAL, MIA, SAS
- For each team, determine the mean, variance, and standard deviation for the following statistics:
  - PTS, AGE, FGA
- Based upon your calculated summary statistics, determine which team had the biggest difference in total season points scored across all of their players.
- Based upon your calculated summary statistics, determine which team had the least variable player age. What was their average player age?
- Based upon your calculated summary statistics, determine which team had the least variability of field goal attempts per player.
- Create a new sheet in the workbook and name the sheet ‘Cleaveland Z-Scores’.
- Within this new sheet, copy over the Player and PTS columns from the raw data for only the CLE team.
- Calculate the z-score for the overall points per player across the whole team.
- Based upon your calculated z-scores, determine which player had the largest difference in total points from the mean of the team.

Suggested Time: 15 minutes



Time’s Up! Let’s Review.



Instructor Demonstration

Quantiles, Outliers, and Boxplots

Real world data can contain extreme values.

01

Some summary statistics, such as the mean, take into account all values of a data set.

02

Extreme values can skew these statistics!

03

Be Careful When Describing Real-World Data

But how can we summarise real-world data?

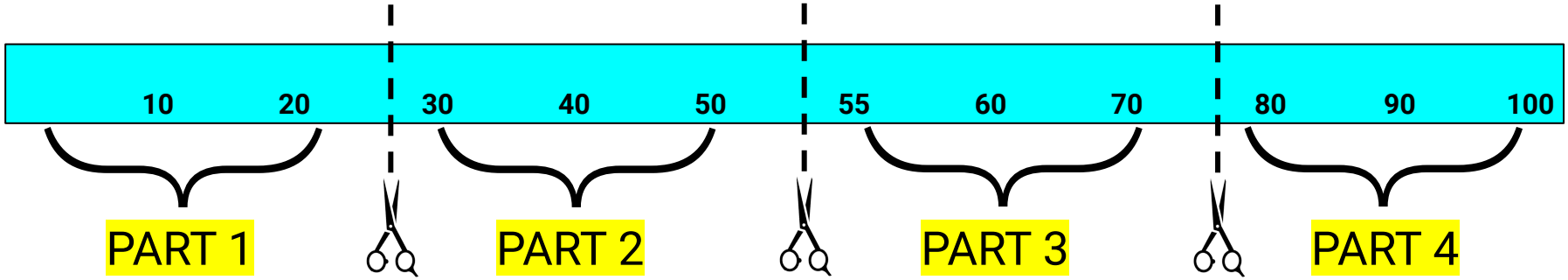
We Can Use Quantiles to Describe Segments of a Data Set!

Quantiles separate a sorted data set into equal-sized fragments

The two most popular types of quantiles are **quartiles** and **percentiles**

01 **Quartiles** divide the data set into four equal parts

02 **Percentiles** divide the data set into 100 equal parts



< Demo Time >

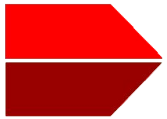
Extreme Values May Not Always Be Reliable

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.



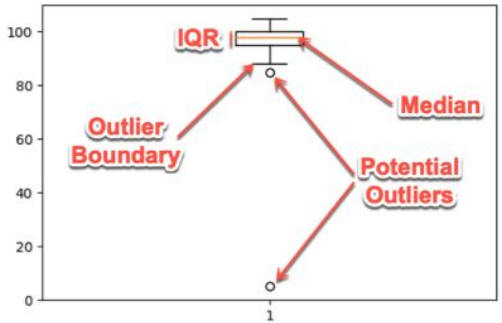
Outliers can inaccurately skew a data set.
Can cause us to misrepresent the actual data

Always be cautious with extreme values.

There Are Two Ways to Identify Potential Outliers

01 Qualitatively

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

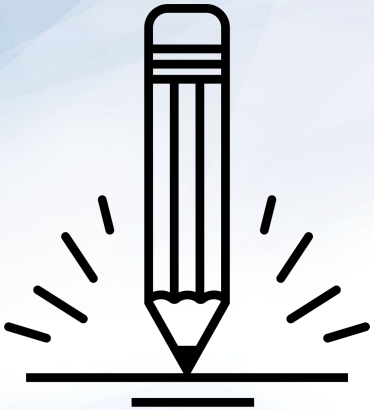


02 Quantitatively

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

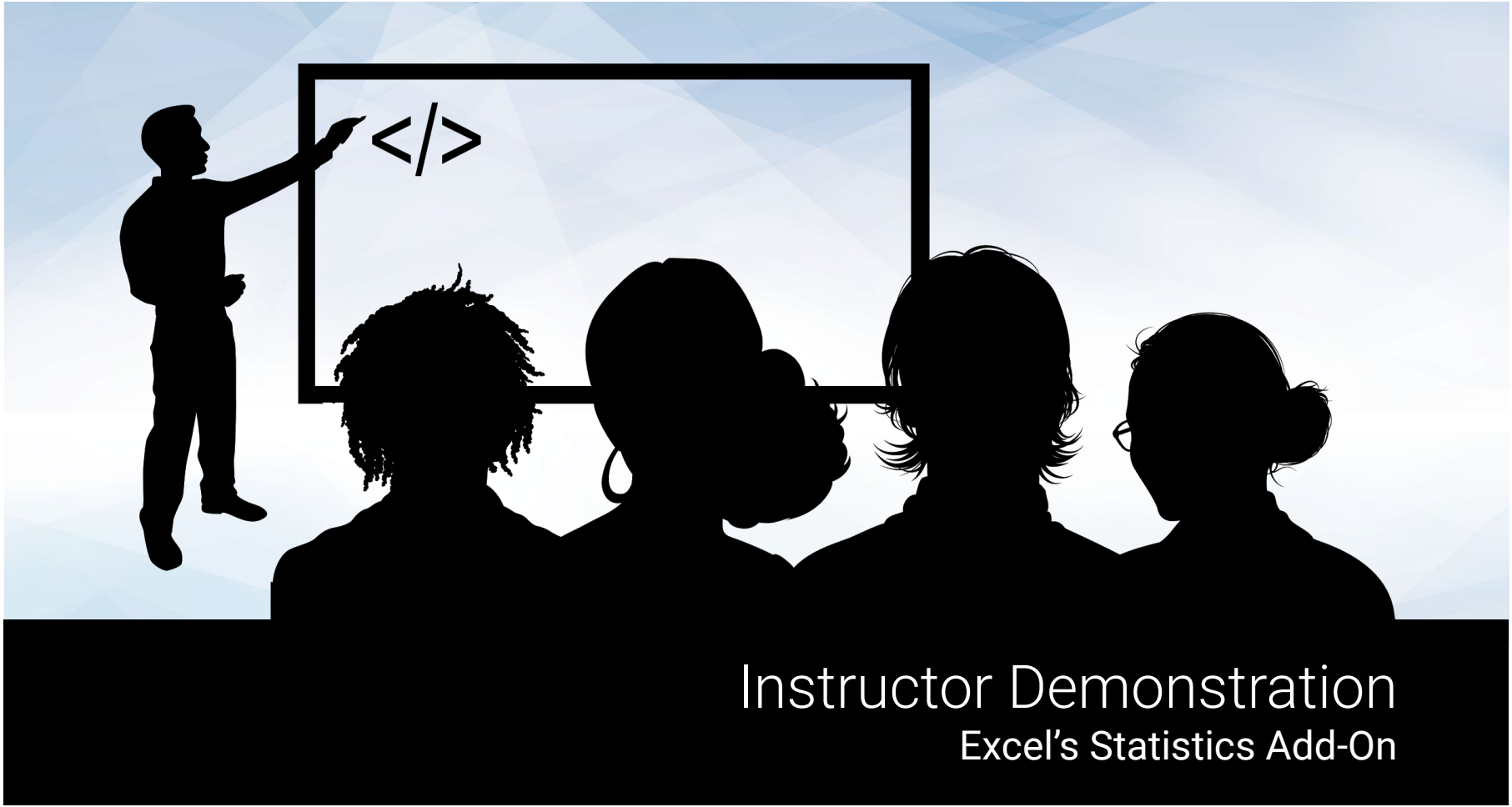
Variance, Standard Deviation and Z-Score Review Instructions

Instructions:

- Open up the activity workbook and familiarise 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

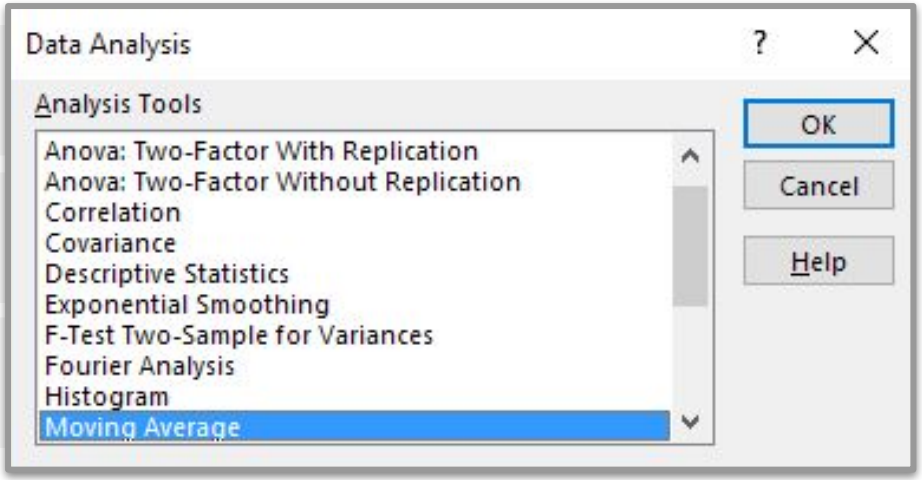
Excel Is a Great Foundational Tool

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, and ANOVA



All of these functions we will cover throughout the course!

Analysis ToolPak Is Not Designed for In-Depth Data Analytics

01

Excel struggles with medium to large data sets

- >200 columns or >100 000 rows
- Depends on machine

02

Excel does not automatically record parameters for statistical tests

03

Excel's Analysis ToolPak **should** be used for

- Gut-checks
- One-off analysis

How to Install and Use the Excel Analysis ToolPak: Mac

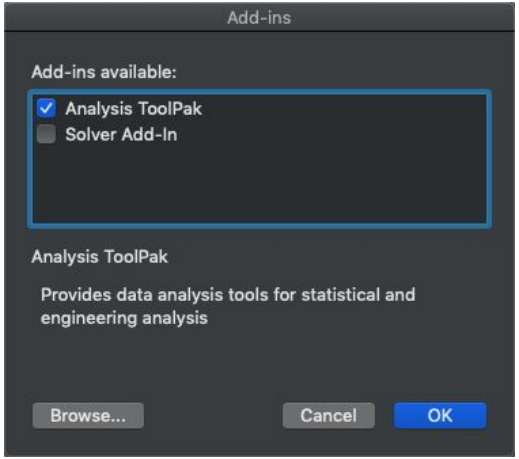
To Install:

01 Go to the ‘Tools’ menu in Excel.

02 Select the ‘Excel Add-Ins’ option.

03 Enable the ‘Analysis ToolPak’ option.

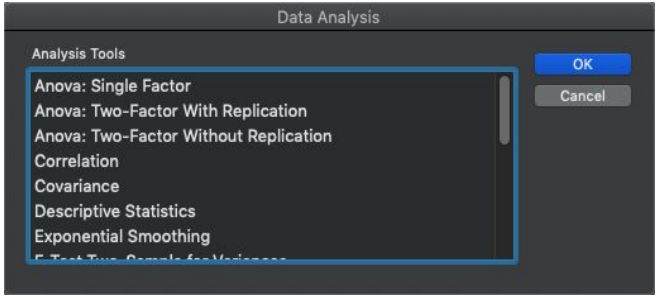
04 Press ‘OK’.



To Use:

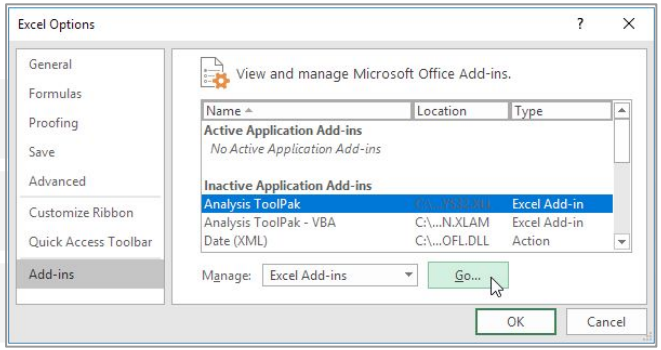
01 Go to the ‘Data’ menu in Excel.

02 Select the ‘Data Analysis’ option.



How to Install and Use the Excel Analysis ToolPak: PC

To Install:



01 Click the ‘File’ tab.

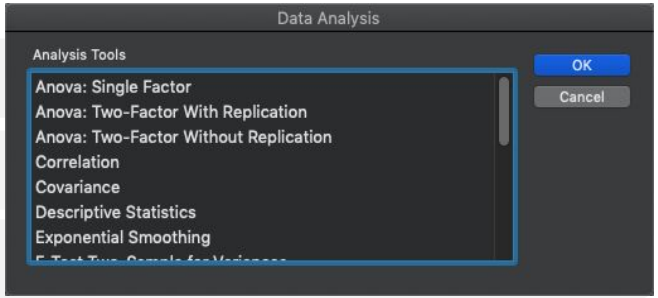
02 Go to ‘Options’.

03 Select the ‘Add-Ins’ category.

04 In the ‘Manage’ box, select ‘Excel Add-Ins’ and click ‘Go’.

05 In the ‘Add-Ins’ box, enable the ‘Analysis ToolPak’ and click ‘OK’.

To Use:



03 Go to the ‘Data’ menu in Excel.

04 Go to the ‘Analyse’ section.

05 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?