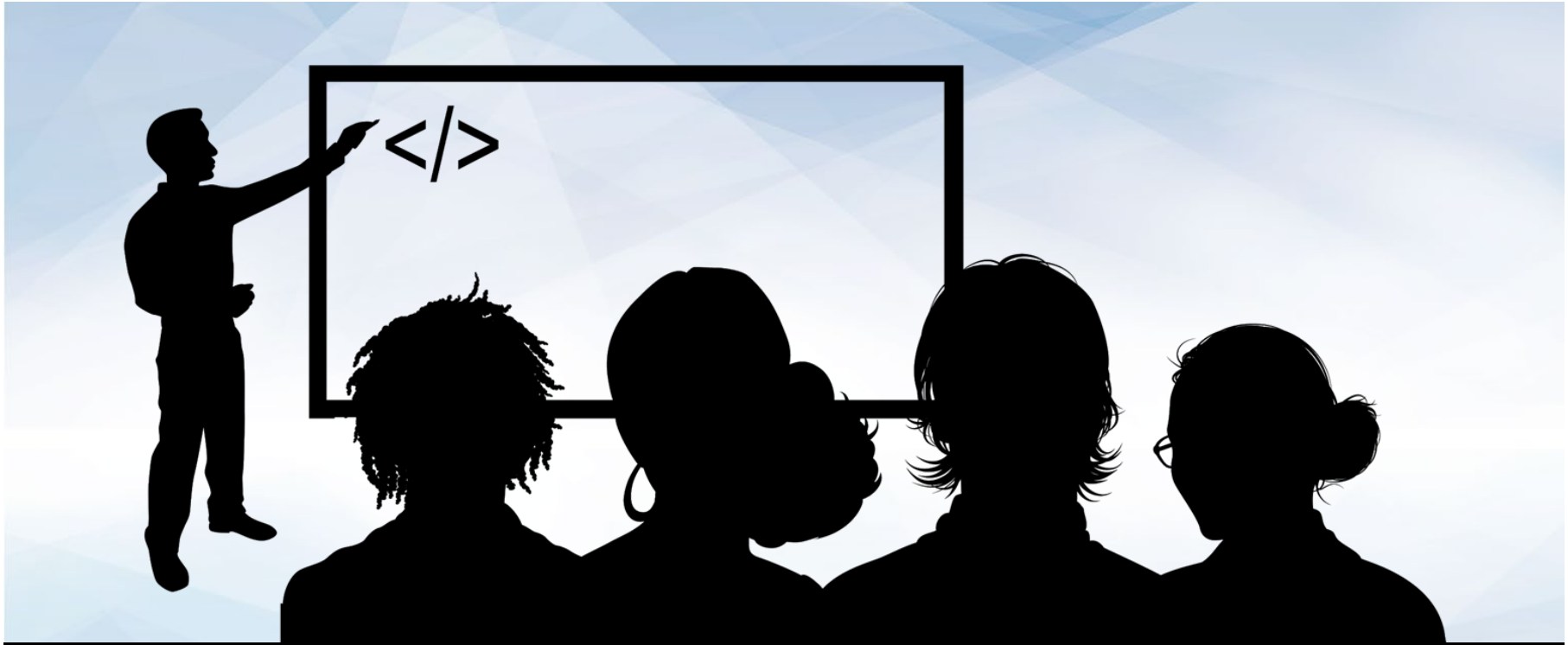
Excel Statistics

Data Boot Camp

Lesson 1.4


WELCOME



Instructor Demonstration

Variance, Standard Deviation and Z-Score

What are the three measures of central tendency?

A

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

- 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

- 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

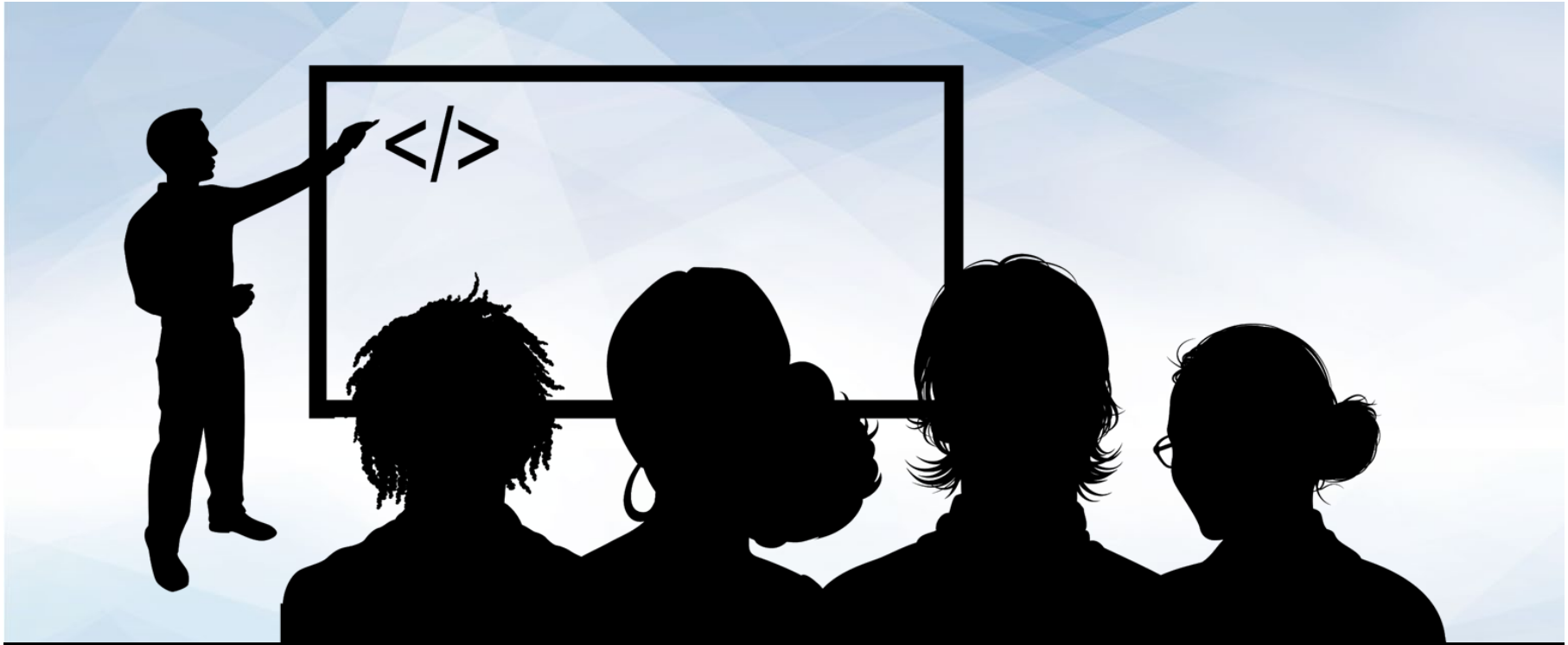
Variance, Standard Deviation and Z-Score Review

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

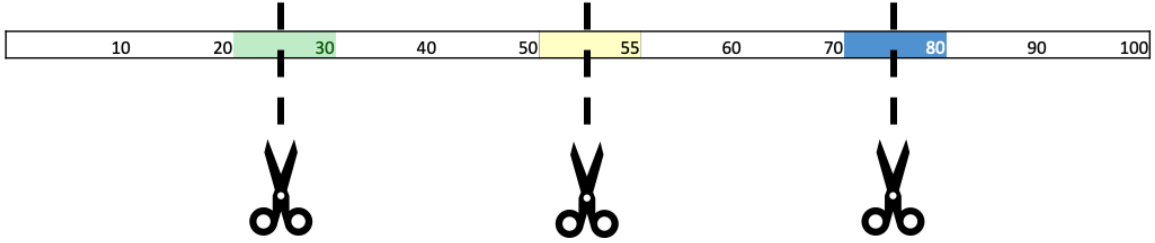
Be careful when describing real-world data

- 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 >

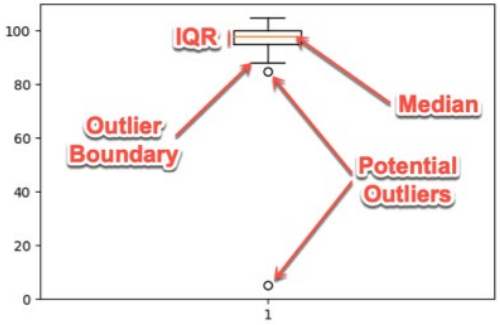
Extreme values may not always be reliable

- **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

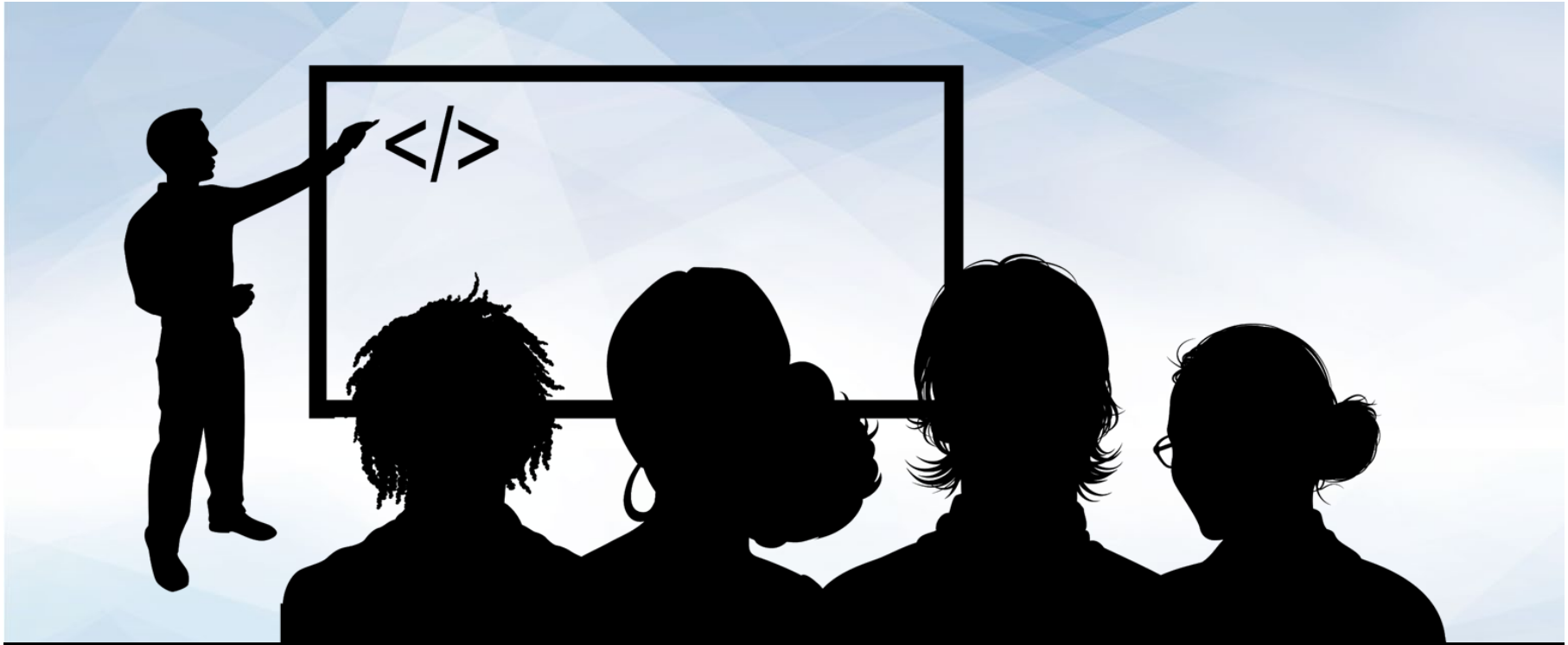
Instructions

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

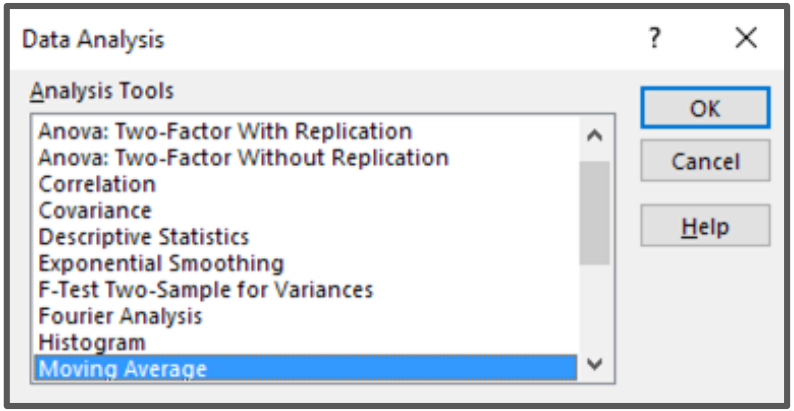
Excel’s Statistics Add-On

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!



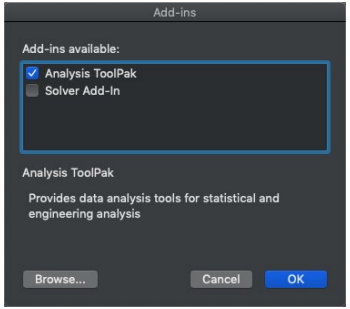
Analysis ToolPak is not designed for in-depth data analytics

- 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

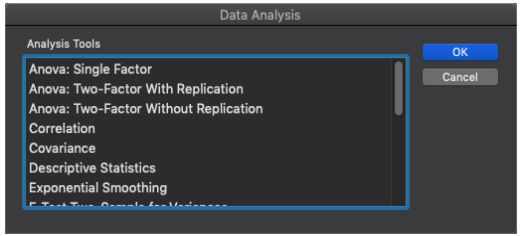
To Install:

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



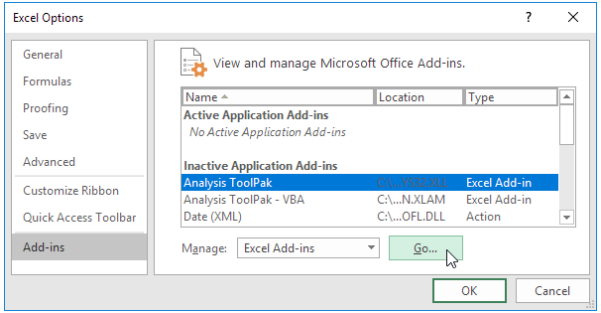
To Use:

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



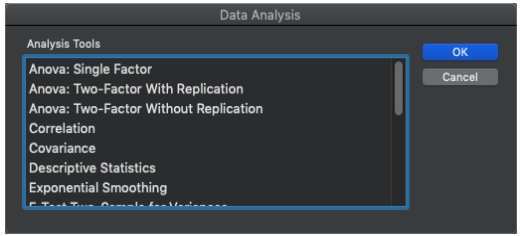
How to install and use the Excel Analysis ToolPak - PC

To Install:



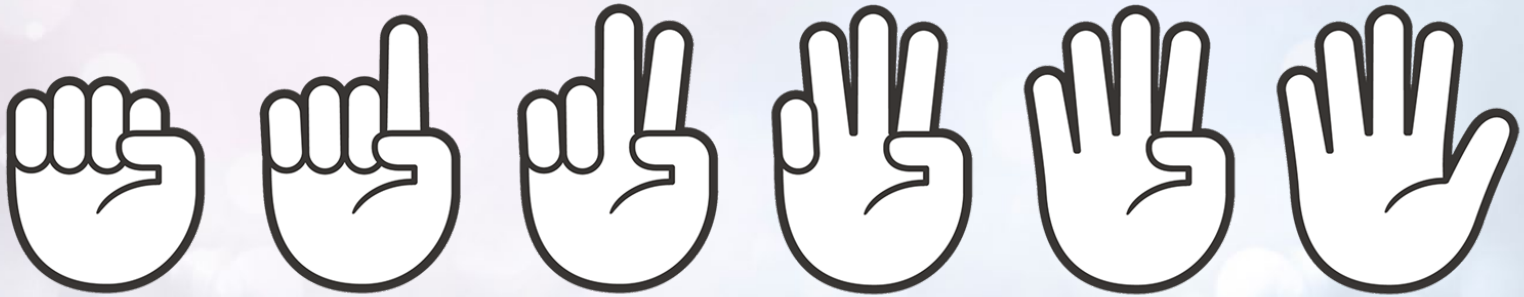
1. Click the File tab
2. Go to Options
3. Select the Add-Ins category
4. In the Manage box, select Excel Add-ins and click Go
5. In the Add-Ins box, enable the Analysis ToolPak and click OK.

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 calculating summary statistics in Excel?