**Data Analytics Lesson 4: Dynamic Data Referencing**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**OBJECTIVES**

* Use data functions [VLOOKUP and HLOOKUP] to manipulate data sets
* Use Pivot tables to provide a fast, flexible way to look at different categorizations and aggregations
* Filter data using report filtering functionality in Excel
* Appropriately select column and rows to pivot on
* Summarize data using the value aggregation functions in Excel
* Create calculated columns within Excel pivot tables in order to create new data value
* Use scatter plot charts to show correlation of data points that may be unclear from data in a table

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
LESSON MAPPING**

**DESCRIPTION:**

In this lesson, we’ll be exploring data and aggregating data with lookups and pivot tables. We’ll be focusing on steps three, four, and five of the analytics workflow: understand the data; prepare, structure, and clean the data; and analyze the data with statistics and visualizations (by looking at Scatter plots).

**STUDENT PRE-WORK**

Students should download the following datasets from the class Schoology;

* AN\_Lesson\_4\_lookup\_exercise\_student.xlsx
* AN\_lesson\_4\_pivot\_table\_exercise\_student.xlsx
* AN\_lesson\_4\_scatter\_plot\_exercise\_student.xlsx
* AN\_lesson\_4\_you\_do\_student.xlsx

**INSTRUCTOR PRE-WORK**

Look through the following:

* [Lesson 4 Deck](https://docs.google.com/a/generalassemb.ly/presentation/d/1MDWuXsjguoEHrAR6OZmNPhLsk6VknOUxSdv2AYU6kFU/edit?usp=sharing)
* [YOU DO Prompt](https://docs.google.com/document/d/1yrXyxKQ9u_SKbUVvNvpSfz11b700DlpjIt_T1cza0RI/edit) to be distributed to students
* [AN Lesson 4 Visualizations Doc](https://drive.google.com/open?id=1kXzTr5O3Io7IQBDWP9QweHmT4nTFi-q1sj_wBOJc9L4)

Find and work through the solutions to all exercises and datasets needed in the spreadsheets below:

* AN\_Lesson\_4\_lookup\_exercise\_instructor.xlsx
* AN\_lesson\_4\_pivot\_table\_exercise\_instructor.xlsx
* AN\_lesson\_4\_scatter\_plot\_exercise\_instructort.xlsx
* AN\_lesson\_4\_you\_do\_instructor.xlsx

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**GUIDE**

|  |  |  |
| --- | --- | --- |
| **Timing** | **Topic** | **One Sentence Description of Section** |
| 10 min | Opening | Review exit tickets; Introduce project 1; Introduce Deloitte context |
| 20 min | Introduction | The importance of data referencing and aggregation |
| 20 mins | WE DO | VLOOKUP, HLOOKUP practice |
| 10 mins | Intro | What are pivot tables? |
| 25 mins | WE DO | Pivot Table practice |
| 15 mins | Intro | Use cases, drawbacks, and demonstrating scatter plots |
| 5 mins | I DO | Creating a scatter plot |
| 55 mins | YO DO | Creating visualizations, aggregations using LOOKUPS and Pivot Tables |
| 10 mins | Closure | Exit tickets + recap of class activity |

**LESSON OUTLINE**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Opening (10 mins)** - Slides 2-9

* Review exit ticket data from previous class
* Select followup questions from previous class’ objectives
* Introduce agenda and objectives for this class
* Remind students of anything project-related: deliverables, timelines, etc.
* Connect this class to the analytics workflow
* Introduce the mock business situations we’ll be using in this class and the corresponding datasets:

You’re working for Deloitte and you’ve been assigned to a project for a client that’s a hedge fund. They’re looking for someone to work with their analysts and provide them with aggregated, summarized data they can use to derive insights from the S&P 500, the WTO, and other country demographics.

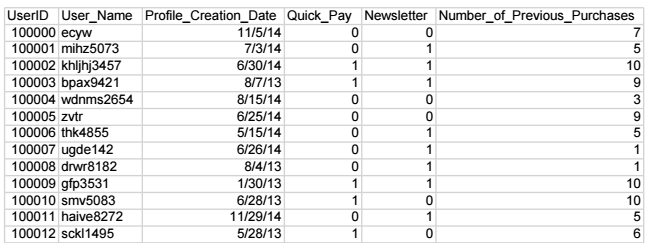
**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Intro - Data Referencing, Data Aggregation, and V-Lookup and H-Lookup (20 mins)** - Slides 11 - 14

# Talking points:

* **Introduce** the importance of data referencing and aggregation:
  + Saved data will be used by managers, statisticians, stakeholders to analyze the workings of software-supported systems and make executive decisions.
  + “Analysis can only take place by methods of aggregation. There's no one in the world who can look at a million rows of raw data and glean insight. The data has to be summed, averaged, standard deviate-ed etc. to make any sense to a human being.” - StackOverflow.com
* **Introduce** V-Lookups and H-Lookups:
  + Easy and effective lookup functions that scan rows and columns for the desired data
    - They are better than normal searches because they are
      * faster
      * easier to change the query
    - Normal searches will just give you the exact cell but with a LOOKUP you can change another cell’s contents depending on what you are looking up
      * It’s a formula as opposed to search
    - V-Lookups (vertical lookups) scan columns
    - H-lookups (horizontal lookups) scan rows
  + They take in a table (usually a subset of the spreadsheet), a column to search in, and a column to output.
    - Basically you enter a value, it searched for that value in column A, and gives you the corresponding value in column B
* Why and when do we use them?
  + Most often used to perform quick searches for a single object in a long column/row
  + **Function:** To run a search on a datatable by a value from the first column
  + **Output:** A value from any cell on the same row from the first column value
  + **Importance**: VLOOKUP and HLOOKUP allow you to connect and lookup information from different datasets.
  + **Explain Syntax**: =V/HLOOKUP(lookup\_value, table\_array, col\_index\_num, [range\_lookup])

**Example:** Imagine that you’re a customer service rep needing to find how many purchases a customer had student, using a VLOOKUP we can do this very quickly. You can create a second sheet and use LOOKUPs across the sheets to bring elements from the first sheet into the other. (To be explained in the next activity).



**Key Idea:** Lookups help you look for values within a range, to find out other associated values with V and H Lookups; you can take larger clean data sets and display specific subsets of data.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**WE DO (students practice with instructor) V-Lookup and H-Lookup (20 mins)** - Slides 16 - 17

**Directions to instructors:**

Be sure students have the AN\_Lesson\_4\_lookup\_exercise\_student.xlsx spreadsheet. Using the business context introduced at the top of the lesson. Fill in the “Assignment” tab with students explaining each step of your process and using the talking points below.

**Directions to students:**

Open the AN\_Lesson\_4\_lookup\_exercise\_student.xlsx to the assignment tab. Our goal is to provide our hedge fund analysts with a summary tab of information regarding the security ticker, name, address, and GICS sector. With the instructor, fill in this tab using VLOOKUP for the top table and HLOOKUP for the bottom table.

**Talking Points:**

* We’re going to be writing more and more powerful formulas as we progress in this class
* If this tab is filled in properly, we should be able to enter any ticker symbol and populate the remainder of the row.
* The goal, as an efficient analyst, is to decrease the amount of work you’re doing - your work should be agile. If you see a task that is repeatable, you should aim to write a powerful formula to do the work for you in the future.
* **Bonus:** Demonstrate fuzzy matches for both v/h lookups, why you would use one vs. the other.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Intro - Pivot Tables (10 mins)** - Slides 19-21

**Talking Points:**

What are pivot tables?

* Aggregation technique for clean data that allows us to chart and visualize certain combinations of data
* A good tool to find relationships between different columns - and subsets of those columns - in a table.
* Classifies numeric data in a list based on other fields in the list.
* **General purpose**:
  + Quickly summarize data from a worksheet or from an external source - calculate totals, averages, counts, etc. based on any numeric fields in your table,
  + Provide the most straightforward way to explore different, multidimensional aggregations in Excel.
  + Generate charts from your pivot tables
  + Find relationships between columns
  + Allow users to explore ore about the datasets they’re working with
  + When we want a way to smart autofilter based on advanced criteria
  + Assuming students know filtering/advanced filtering, pivot tables can be used in conjunction to:
    - Explore and hide extraneous data using table views, Autofilter, and criterion range.
    - Take larger clean data sets and display specific subsets of data and pivot on it

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**WE DO - Pivot Tables (25 mins)** - Slides 23 - 25

**Direction to instructors:**

Provide and introduce the context below. Be sure to allow students download the dataset - AN\_lesson\_4\_pivot\_table\_exercise\_student.xlsx) - to work in groups to determine what questions can be answered from the data first, and then review and go through the questions we’ll be looking to answer with our work building the pivot tables:

**Overall trend questions:**

* Which countries with the highest import growth in the last 5 years?
* Which have the lowest (or negative) growth?
* Which import indicators have the highest volume and which have the highest growth over the time of the data?

**Using toggle by country table:**

* What were Clothing Exports In Germany in 2007?
* What were total Imports in the the US in 2009?

**Using toggle by sector table:**

* Who were the three largest total exporters in the data in 2013?
* Who imported more fuel in 2011, the US or China?

Then, with the students following along, create a summary/analysis she using separate pivot tables that answer the questions and create the pivot tables in such a way that a user can explore them to get further results.

Review what information you would include in a brief statement (a paragraph or two) answering the above questions and noting anything else interesting you discovered to your supervisors.

**Directions to students:**

The hedge fund you’re assigned to is considering expanding into other countries. Again, the analysts you’re working with have asked you to provide aggregated, summarized data on expansion of imports and exports of countries in the world. Given the World Trade Org (WTO) data, create two summary tables that highlights the breakdown of Imports & Exports by sector. The analysts want to be able to toggle by country and by sector.

You have located economic data on imports and exports from the WTO website <http://stat.wto.org/StatisticalProgram/WSDBStatProgramHome.aspx?Language=E> (we already have the workbook - AN\_lesson\_4\_pivot\_table\_exercise\_student.xlsx) You have downloaded several years of information and you have saved it in the World\_trade\_org\_Import\_export\_co\_e.csv file.

Before diving into aggregation, take two minutes to identify, with a partner, the business needs, and what questions you might be able to answer. After we go over the important questions we’ll be able to answer, we’re going to create two summary charts using separate pivot tables that answer the questions and create the pivot tables in such a way that a user can explore them to get further results. Then we’ll brainstorm what we could send in an email to our analysts to give them a recap.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Intro - Scatter Plot Chart (15 mins)** - Slides 27-37

# Talking Points:

* Define Scatter Plot Chart:
  + A scatter plot is a chart obtained by mapping numeric values to a pair of orthogonal axes.
  + Dots represent a data point in the data set and its position is given by the x and y position obtained by using the values of the two attributes that are mapped to the two scatter plot axes.
  + **Example**: [Figure](https://docs.google.com/document/d/1kXzTr5O3Io7IQBDWP9QweHmT4nTFi-q1sj_wBOJc9L4/edit) [1](http://#) - food dataset mapping each food on a “sugar” vs. “calories” axis pair; Each dot represents one food and as you can see there is an association between sugar amount and calories, but not for all foods.
    - **Explain:** if there appears to be an angled line pattern [(see figure 2)](https://docs.google.com/document/d/1kXzTr5O3Io7IQBDWP9QweHmT4nTFi-q1sj_wBOJc9L4/edit) that appears in a scatterplot, it signifies a correlation between the the x-axis and y-axis values; the closer the line isto 45 degrees, the stronger the correlation.
* Identify Scatter Plot Variants:

1. **Bubble Chart**: same as scatter plot but a third numeric attribute is mapped to the size of the bubble.
   * + **Example**: [Figure 3](https://docs.google.com/document/d/1kXzTr5O3Io7IQBDWP9QweHmT4nTFi-q1sj_wBOJc9L4/edit), same scatter plot as above but now the size of each circle is scaled according to protein amount (the third attribute mapped in the scatter plot).

2. **Scatter Plot with Colored Dots**: a scatter plot can also be enhanced by mapping the color of the dots to an additional attribute (the same way we just did with size). Color coding can be used both to map a categorical as well as a numeric attribute. Typically, categorical attributes work best on this case. In [figure 4](https://docs.google.com/document/d/1kXzTr5O3Io7IQBDWP9QweHmT4nTFi-q1sj_wBOJc9L4/edit), we can see the kcal content additionally classified by food type.

3. **A combination of bubble and colored dots:** In [figure 5](https://docs.google.com/document/d/1kXzTr5O3Io7IQBDWP9QweHmT4nTFi-q1sj_wBOJc9L4/edit), you can see how we, again, map size with color at the same time, effectively mapping a total of 4 attributes in one plot.

* + - **Note:** When data is time-oriented, that is, the values depicted in the plot change with time, “small multiples” can be used to see how the configuration changes. Similarly a scatter plot can be animated to represent change in time, even though this solution is not always effective, especially if too many things change at the same time.
* **Describe** use cases for scatter plots:
  + to show the association between 2, 3 or 4 variables.
* **Explain** drawbacks:
  + The biggest issue with scatter plot is data size: if there are too many data points it can become too cluttered to show useful information.
  + Mapping color to an attribute with too many categories (hard to distinguish more than 5 or 6 colors).
  + Another similar problem is when the data is highly skewed the dots may all be concentrated in a small area.
* **Explain** helpful hints using [figures 6 and 7](https://docs.google.com/document/d/1kXzTr5O3Io7IQBDWP9QweHmT4nTFi-q1sj_wBOJc9L4/edit):
  + transparency can help to some extent
  + use density plots rather than scatter plots; a density plot subdivides the area in bins and colors each bin with a color intensity proportional to the number of data items that fall in the bin.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**I DO - Creating a scatterplot in Excel (5 mins) -** Slide 39

**Directions to instructors:**

* **Provide** context:
  + Looking to get more information, the analysts you’re working with are now asking for information about the development of the countries you’ve gathered information; they are specifically looking to know more about the country’s demographics.
  + Given the World Health Organization, in the following 2 tabs, create a scatter plot that demonstrates the relationship between Fertility & Life Expectancy globally.
* **Demonstrate** creating a scatter plot in excel using the AN\_lesson\_4\_scatter\_plot\_exercise\_student.xlsx workbook
* **Explain** the conclusions that can be easily derived from this visualization
  + Look at three different years: 1960, 1980 & 2000. What do you think can be concluded, if anything given this data?

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**YOU DO - LOOKUPS, Pivots and Visualizations to Analyze Data (55 mins)** - Slide 41

**Direction to instructors:**

Have students download the AN\_lesson\_4\_you\_do\_student.xlsx Excel workbook and provide the context below below.

**Directions to students:**

You’ve gotten an email from the analyst you’ve been aggregating data for. Read the email and use the workbook they sent over to create the visualizations and the aggregations they’re looking for. Finally answer any questions they have using your work.

**The email:**

“I need to do some analytics on the stocks that we are looking at for a new portfolio option we’re providing to our clients. I’ve attached a spreadsheet I’d like you to fill in for me. In the tab labeled ‘Raw Price Data’, there is stock price history information on the stocks that we are considering. The stocks are labeled by the codes given by the third party vendor. I've attached a key with each code and the corresponding ticker in the tab ‘Key’.

Also I've attached info about the S&P500 on the tab called ‘S&P500’ so that you can figure out the names, sectors, etc. for each security.

I'm trying to clean up the data and put it a format so that I can see all the prices by date, side by side. Take a look at the deliverable in the assignment tab for what I'm looking for.

After that, please use the data to just show the price at the end of each calendar year. I've already started it for you, you just need to fill in the data on the Summary Table Tab

That data will flow through another table that will calculate the yearly returns for each stock.

I want you to make four scatter plots based on the yearly return data.

1) Comaring the IT stocks vs the S&P 500 yearly returns

2) Comaring the financials stocks vs the S&P 500 yearly returns

3) Comaring the energy stocks vs the S&P 500 yearly returns

4) Comaring the industrials stocks vs the S&P 500 yearly returns

Tell me if you think there is a strong relationship between the S&P500 yearly returns and the each sectors yearly returns.

Use a scatter plot to visualize the four sectors!

Thank you!”

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**CLOSING DISCUSSION (10 minutes)**

**Directions to instructor:**

Review answers to final YOU DO

Distribute link for exit ticket