GENERAL ASSEMBLY

DATA AGGREGATION IN EXCEL

Matthew Morris

Costco Wholesale

LEARNING OBJECTIVES

- Summarize data using pivot tables.
- Use Excel aggregation commands, MIN, MAX, SUM, AVERAGE, COUNT, and their conditional variants, COUNTIF, COUNTA, COUNTIFS, COUNTBLANKS, to summarize data sets.

OPENING

OPENING

- Now we are going to learn about aggregate functions. These functions *summarize* our data in various ways.
- Aggregate functions are helpful on their own, but they are also a large part of *pivot tables*.
- Pivot tables are tremendously helpful when creating summaries of data, and we will learn how to create them today.

INDEPENDENT PRACTICE: SCAVENGER HUNT

ACTIVITY: SCAVENGER HUNT

EXERCISE

DIRECTIONS

1. Open L3_scavenger_hunt.xlsx
2. Complete the scavenger hunt activity (15 min)

DELIVERABLE

Completed scavenger hunt

GUIDED PRACTICE: HOW TO USE AGGREGATE FUNCTIONS

WHAT ARE AGGREGATE FUNCTIONS?

- Aggregate functions allow you to *summarize* information using formulas in Excel.
- Let’s learn about Excel’s aggregating functions and apply them to our scavenger hunt. We’ll be able to check our scavenger hunt answers with the new functions we are about to learn.

=MIN(...)

- MIN: Finds the minimum value of a range of numbers.
  - Syntax:

    =MIN(number1, [number2], ...)

  - Solution to our scavenger hunt:

    =MIN('2014_acs_select'!B:B)

=MAX(...)

- MAX: Finds the maximum value of a range of numbers.
  - Syntax:

    =MAX(number1, [number2], ...)

  - Solution to our scavenger hunt:

    =MAX('2014_acs_select'!O:O)

=SUM(...)

- SUM: Finds the sum of a range of numbers.
  - Syntax:

    =SUM(number1,[number2],...)

  - Solution to our scavenger hunt:

    =SUM('2014_acs_select'!E:E)

=AVERAGE(...)

- AVERAGE: Finds the average of a range of numbers.
  - Syntax:

    =AVERAGE(number1, [number2], ...)

  - Solution to our scavenger hunt:

    =AVERAGE('2014_acs_select'!AI:AI)

=COUNT(...)

- COUNT: Counts the number of *numeric* values in a range.
  - Syntax:

    =COUNT(value1, [value2], ...)

  - Solution to our scavenger hunt:

    =COUNT('2014_acs_select'!D:D)

=COUNTIF(...)

- COUNTIF: Counts the number of values in the range that meet the given criteria.

  - Syntax:

    =COUNTIF(range, criteria)

  - Solution to our scavenger hunt:

    Part A: =COUNTIF('2014_acs_select'!B:B,"<25")

    Part B: =COUNTIF('2014_acs_select'!AL:AL,"=High")

    The criteria has to be in double-quotes (“…”)

=COUNTA(...)

- COUNTA: Counts the number of non-blank cells, not just the number of numeric cells.

  - Syntax:

    =COUNTA(value1, [value2], ...)

  - Solution to our scavenger hunt:

    =COUNTA('2014_acs_select'!A:A)-1

    Note: We have to subtract one to avoid the header row!

=COUNTIFS(...)

- COUNTIFS: Similar to COUNTIF, except it can take many ranges with many criteria.
  - Syntax:

    =COUNTIFS(criteria_range1, criteria1, [criteria_range2, criteria2]…)

  - Solution to our scavenger hunt:

    =COUNTIFS('2014_acs_select'!C:C,"<35",'2014_acs_select'!D:D,">=35")

=COUNTBLANKS(...)

- COUNTBLANKS: Counts the number of blank cells in the range. This is the complement to COUNTA.

  - Syntax:

    =COUNTBLANK(range)

  - Solution to our scavenger hunt:

    =COUNTBLANK('2014_acs_select'!A1:AM1446)

INDEPENDENT PRACTICE: USING AGGREGATE FUNCTIONS

ACTIVITY: USING AGGREGATE FUNCTIONS

DIRECTIONS

1. Open L3_independent_activity_p1.xlsx
2. Based on your experience, choose either the BASE or STRETCH tab to complete (20 min)

You may work with a partner, checking in with each other after answering each question.

DELIVERABLE

Complete BASE or STRETCH tab in L3_independent_activity_p1.xlsx

INTRODUCTION: PIVOT TABLES

INTRODUCTION TO PIVOT TABLES

- Pivot tables take aggregation and filtering to a whole new level. They allow you to quickly create aggregations, slices, filters, and more using your data as a source, and all without having to write your own formulas.
- Pivot tables are a massive topic, but we will learn enough today to use them efficiently. Entire books have been written on the topic.

DEMO: COMPONENTS OF A PIVOT TABLE

THE DIAMONDS DATASET

- To demonstrate pivot tables, let's consider this simple dataset. This is known as the diamonds dataset, and is often used to provide examples of how to use the R programming language. In fact, it comes with the R programming language, so everyone can use the same dataset to easily demonstrate code.

THE DIAMONDS DATASET

The diamonds dataset has 10 variables:

- Price: price in US dollars ($326-$18,823)
- Carat: weight of the diamond (0.2-5.01)
- Cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- Color: diamond color, from J (worst) to D (best)
- Clarity: a measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best))
- x: length in mm (0-10.74)
- y: width in mm (0-58.9)
- z: depth in mm (0-31.8)
- Depth: total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43-79)

CREATE A PIVOT TABLE WITH DIAMONDS DATASET

- Create a pivot table by clicking PivotTable on the Insert ribbon.
  - Make sure you have a cell selected in the diamond data to pre-populate the range.
- Verify that the Table/Range auto-selected is the diamonds data. If not, you can change it before moving on.
- Decide where you want to put the pivot table. Let's create a new worksheet for this example (default).
- Click OK.

COMPONENTS OF A PIVOT TABLE

Pivot tables have four primary components:

- Filters
  - What data should we include *at all* in our pivot table? We are first able to filter out data we do not want to include.
- Rows
  - What *unique data values* do we want to have as rows in our table?
    - These values must exist in the data. If they don't, you need to create them first.

COMPONENTS OF A PIVOT TABLE

Pivot tables have four primary components:

- Columns
  - What *unique data values* do we want to have as columns in our table?
  - These values must already exist.

- Values
  - These are the values that will be in the cells of our table. To create them, we will have to tell Excel how we want them aggregated, using aggregation functions we have already learned.

DRAG AND DROP VARIABLES TO CREATE PIVOT TABLE

Let’s demonstrate with a few examples to see how this works.

- Average price per color:
  - Drag “color” to Rows.
  - Drag “price” to Values.
  - To fix the average price: click “i” next to “Sum of price” in Values.
  - Change “Summarize by” to “Average”.
  - This will properly take the average, but let’s make sure our formatting is correct. These averages are dollar amounts, so click “Number…” and format as currency.

DRAG AND DROP VARIABLES TO CREATE PIVOT TABLE

Let's demonstrate with a few examples to see how this works.

- Average price per color and cut:
  - If we want to calculate average price *per color and per cut*, drag "cut" into Rows.
  - For every color, we now see the price per that color and each cut.
  - What if we wanted cut and then color? Change the order and see the results.
  - This is great, but it would be much easier to parse and compare this data if the colors were down the rows and the cuts were across the columns…
  - Move "cut" to the Columns.

DRAG AND DROP VARIABLES TO CREATE PIVOT TABLE

Let's demonstrate with a few examples to see how this works.

- Average price per color and cut for the top three best clarities:
  - For this, the structure of our pivot table is going to remain the same, but we need to *filter* some of the data out of the calculations.
  - Move "clarity" to Filter.
  - Notice that a drop-down for clarity has been added above the pivot table. Use the drop-down to filter out only VVS1, VVS2, and IF.

DRAG AND DROP VARIABLES TO CREATE PIVOT TABLE

Let's demonstrate with a few examples to see how this works.

- Question about our dataset itself: what is the number of diamonds by cut, color, and clarity?
  - Move cut, color, and clarity to Rows.
  - Change Average of price to Count of price. Reformat to a number.
    - When doing a Count, the variable used for the Values cell can often be any of them… as long as they have values of some kind. Think of it as counting the number of values for a given column. Since the table is a rectangle, this also gives you the number of observations.

PIVOT TABLES

- Pivot tables work very well with this dataset of diamonds because the data is not aggregated. Each row defines only one observation (one diamond), and contains many *categorical* variables. Categorical variables are especially important for pivot tables, because a pivot table will always turn a variable into a categorical variable when placed as a Row or Column variable.

PIVOT TABLES

- Our ACS dataset is different. It already contains data aggregated by census tract. (If it were not aggregated, it would be a set of people, not census tracts.) Because of this, we need to keep the following things in mind to make our pivot tables accurate and streamlined:
- For any values that we want in the rows or the columns, they should be discrete categorical variables.
    - In the case of our ACS data, this often means we will need to create categorical data, as we did with density in the last lesson.

GUIDED PRACTICE: PIVOT TABLES WITH ACS DATASET

PIVOT TABLES WITH ACS DATA

- Example 2: What are the minimum, average, and maximum sizes (in sq. km.) for census groups by density?
  - Rows: density_group
  - Values:
    - Min of area_sqkm
    - Average of area_sqkm
    - Max of area_sqkm

PIVOT TABLES WITH ACS DATA

- Example 3: What is the percentage of each county that lives within each type of density group?
  - Rows: county
  - Columns: density_group
  - Values: Total population
    - Set “Show Data As” to “% of Row”

INDEPENDENT PRACTICE: USING PIVOT TABLES

ACTIVITY: USING AGGREGATE FUNCTIONS

EXERCISE

DIRECTIONS

1. Open L3_independent_activity_p2.xlsx
2. Based on your experience, choose either the BASE or STRETCH tab to complete (20 min)

You may work with a partner, checking in with each other after answering each question.

DELIVERABLE

Complete BASE or STRETCH tab in L3_independent_activity_p2.xlsx

CONCLUSION

CONCLUSION

- We learned about aggregate functions and how they can summarize data.
- We also learned about pivot tables and how useful they can be.
  - They can save so much time and help create aggregations and slices of data that would have been difficult to see before.
  - Understanding the logic of a pivot table will help in both the SQL unit (when using aggregate functions and GROUP BY clauses) as well as Tableau.

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET

CREDITS

CITATIONS

- Diamonds dataset from ggplot2 R package: http://docs.ggplot2.org/0.9.3.1/diamonds.html

RESOURCES

- The Excel documentation on the Microsoft website is extremely helpful when looking up the arguments and behavior of Excel functions: https://support.office.com/en-us/excel