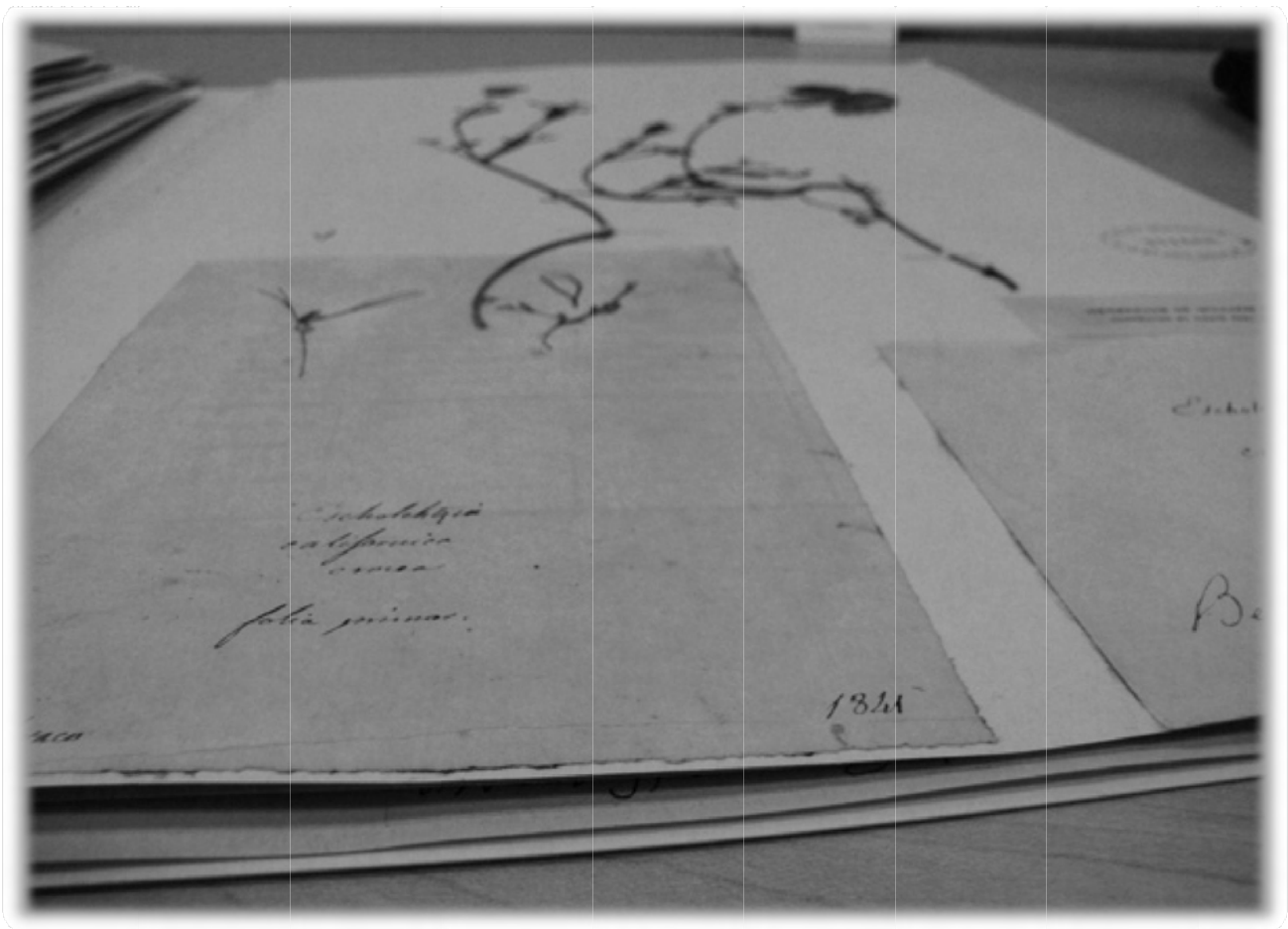


SKELETONS IN THE CLOSET

PRESERVED PLANTS REVEAL PHENOLOGICAL RESPONSES TO CLIMATE CHANGE

This exercise will guide you through the basic processes of exploring long-term phenological data sets. Using a data set derived from herbarium specimens collected from 1906-2009, you'll be guided step-by-step through the processes of organizing, summarizing, visualizing, and analyzing the data using Microsoft Excel. Discussion questions and suggestions for continued learning are included for each section. For more background on herbaria and how they've been used to study phenology, read our *Primer on herbarium-based phenological research*, available on the Education section of the California Phenology Project website (www.usanpn.org/cpp/education) or the USA National Phenology Network (www.usanpn.org/education).



BRIAN HAGGERTY, ALISA HOVE, AND SUSAN MAZER

UNIVERSITY OF CALIFORNIA, SANTA BARBARA

Funding for the development of these materials was provided by the US Geological Survey and the USA National Phenology Network



More phenology education materials and activities are available online, including a *Primer for herbarium-based phenological research*; annotated lectures for universities and the public; guides to establishing and using phenology gardens; seminar modules for undergraduate/graduate students; and standards-aligned K-12 lesson plans. To learn more and to download materials, visit the Education section of the California Phenology Project website (www.usanpn.org/cpp/education) or the USA National Phenology Network (www.usanpn.org/education).

Skills gained in Microsoft Excel by completing this exercise

- Organizing data – sorting and filtering
- Summarizing data – creating pivot tables and pivot charts
- Visualizing data – creating and formatting histograms
- Analyzing data – creating scatterplots and conducting regression analysis

Materials & knowledge needed to participate in this activity

- Microsoft Excel 2007 for PC – Although all available versions of Excel have the functions needed to complete this activity, their navigation, and in some cases their functionality, may be different that what is shown and described here.
- Internet access to download the data set
- Basic knowledge of computer functions

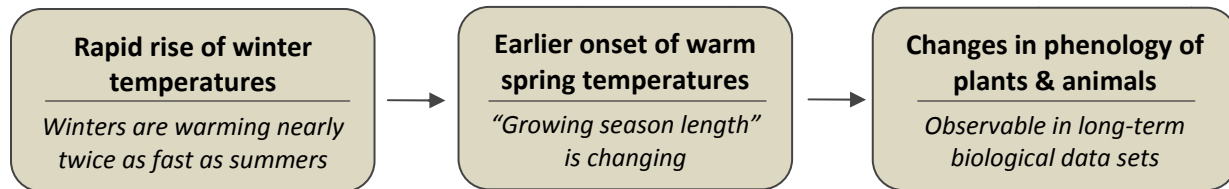


ACTIVITY INTRODUCTION

Our understanding of the global climate system has advanced remarkably over recent decades, along with our awareness that changes in the climate system have affected plant and animal activities. Biologists have been able to detect some of the impacts of climate change on living systems (including both wild and cultivated populations, species, and communities) by exploring data sets from long-term monitoring efforts such as annual surveys of wildflower blooms, land surface “green-up”, animal migrations, insect emergence, and crop harvests.

In a landmark 2007 report, the Intergovernmental Panel on Climate Change (IPCC)¹ summarized 28,671 long-term data series from around the world that document significant biological impacts of climate change. They concluded, among other things, that “Phenology – the timing of seasonal activities of animals and plants – is perhaps the simplest process in which to track changes in the ecology of species in response to climate change.” Although phenological responses to climate change (and the processes that cause them) vary among species and geographic regions, the general process is as follows:

¹ Formed in 1988 by the World Meteorological Organization and the United Nations Environmental Programme, the IPCC is the international body of scientists assessing the available scientific, technical, and socio-economic information relevant to understanding climate change, its impacts, and options for mitigating its consequences on wild and managed systems.



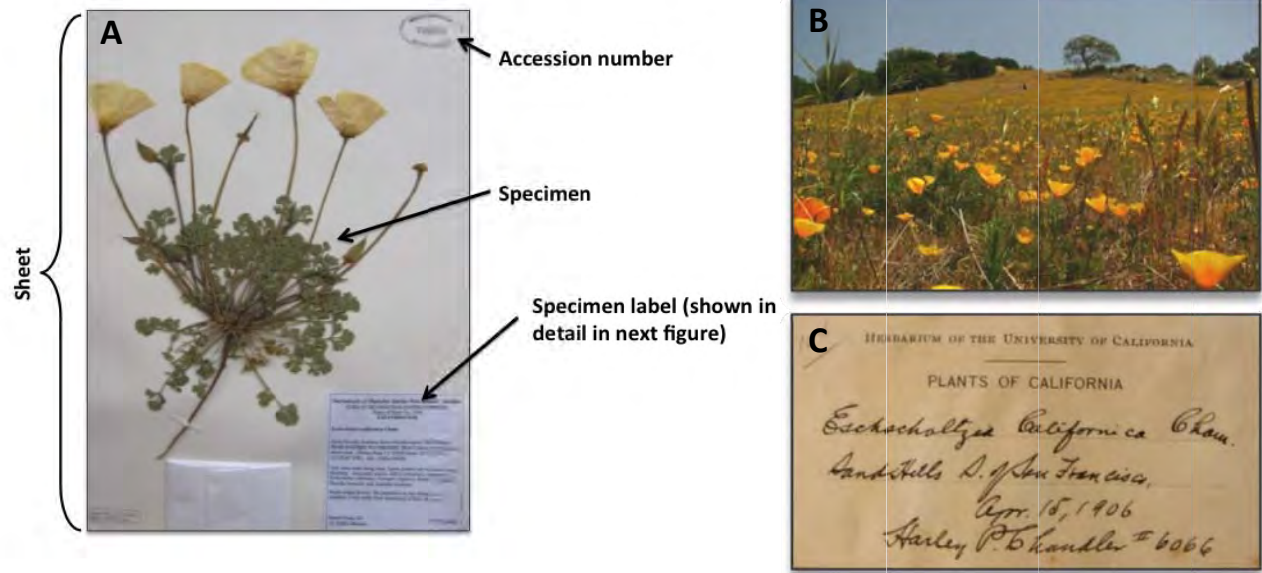
Our understanding of phenological responses to climate change is limited by the availability and scope of long-term data sets. For example, only 3.4% (or about 1,000) of the long-term data series that the IPCC summarized were generated in North America. While impressive progress has been made in the U.S. in recent years to establish a nationwide long-term biological monitoring program for academic, government, and citizen scientists (the USA National Phenology Network – www.usanpn.org), researchers have begun examining other sources of long-term historical data that may document phenological changes in plants and animals. These research projects also have helped to improve the historical context in which phenological monitoring programs are being established. One of the most powerful and extensive data sources available is herbarium records.

An herbarium is a research and education center much like a library or a museum, but where plants collected from wild or cultivated populations are curated and stored in a stable environment. After plants are collected, pressed, and dried, they are mounted on museum-quality paper with an information card (including the when, where, who, and other information pertaining to the collection location, date, and activity). The specimen (or “sheet”) is then assigned an accession number and stored with others in a systematic manner in cabinets. A sheet and a label are detailed on the next page.

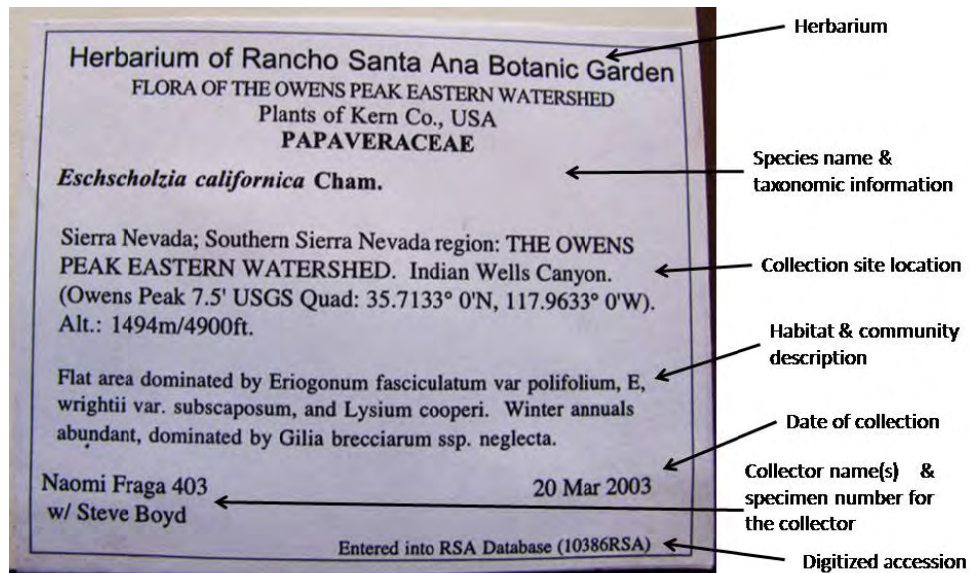
There are over 600 herbaria in the U.S. containing over 260 million specimens. Most herbaria are located at universities and botanic gardens but they may also be found at National Parks and in private collections. In the absence of long-term monitoring data obtained at a single location for many years using standardized methods for observing and recording phenological events or phases, herbaria are perhaps our best repositories of historical information that can be used to study biological responses to environmental changes, including climate change.

For more information on herbaria and how they’ve been used to study phenology and climate change, read our *Primer on herbarium-based phenological research*, available on the Education section of the California Phenology Project website (www.usanpn.org/cpp/education) or the USA National Phenology Network (www.usanpn.org/education).





- A) Example of a preserved and cataloged California poppy herbarium specimen. On this sheet, the entire plant is available for study, including its root and all reproductive structures. The accession number is stamped in the upper-right corner of the herbarium sheet, and a label including information about the collection appears in the lower-right corner. A hand-made envelope (visible at the bottom of the sheet) is often glued to a sheet in order to store any fragments or seeds that become separated from the plant.
- B) Recent hillside population of California poppy in the southern Sierra Nevada Mountains.
- C) Handwritten collection label from a specimen collected 1906. Compare the amount of information provided on this label compared to the label below.



The format of herbarium sheet labels may differ among herbaria, collectors, and time periods. Most labels will include much of the information shown here, with some modern labels having GPS coordinates for the collection site as well (as shown here).

Background on the data set –The data set for this activity comes from a research project underway at the University of California, Santa Barbara designed to assess the effects of climate change on plant phenology in California. The focus of the data set used in this exercise is the widely-distributed and frequently-collected California poppy (*Eschscholzia californica*). Several herbaria were visited (and loans were requested from others through the Consortium of California Herbaria) to examine specimens and to answer the following research question: “Has the flowering time of California poppy changed in California in recent decades?” Based on the reasoning above, we predicted that, in response to a warming climate, we’d find that California Poppies are flowering earlier now than at the beginning of the 20th century. You will be able to answer this question and test this prediction by completing this activity while learning (or reinforcing) basic skills in data processing and analysis.

Data collection – We browsed the herbarium collections to find California poppy plants that were collected shortly after producing their first flowers – the preserved plant had only one or two flowers open, many flower buds, and no wilted flowers or developing fruits. We also examined some plants that were collected at peak flowering (when stems bear a large proportion of flowers relative to flower buds and developing fruits). Only specimens that did not have any missing parts were included in the data set – the entire plant had to be visible for examination. This criterion assured that our assessment of an individual plant’s flowering phase was correct; for example, the flowering status of “missing” branches may have altered our conclusion as to whether a plant was producing its very first flower). We recorded the collection information from each of the specimens into an Excel spreadsheet with the following column headers:

- herbarium
- accession number
- collector name(s)
- date of collection (month, day, year)
- location of collection (county, and elevation, latitude, and longitude if available)

In some cases, more than one individual plant was mounted on one sheet. In these instances, the collector simply collected several plants from the same location on the same day and mounted them on the same sheet. In the data set, therefore, it will appear as though some rows are repeated... in fact, this is due to the fact that several individuals per sheet were included in the study.

Data processing – After visiting each herbarium, we converted the collection date to a “day of year” value. Thus, specimens collected January 1 would be assigned a value of 1, and specimens collected December 31 would be assigned a value of 365 (*leap years are a critical part of this conversion! We identified the leap years in the data set and adjusted the day of year values accordingly*). We also converted all elevation values to meters. In this activity you will perform several other data-processing steps for organizing and exploring the data.

Please note that other types of information were recorded for this research project, and other types of data-processing steps (not covered in this activity) have been taken to analyze the data. Since this is an ongoing research project, however, only a subset of the entire data set is currently included for this activity. Once the research project is complete, the entire data set will be made public through the USA National Phenology Network and this activity guide will be updated.

Data conventions & definitions –

It is general practice among researchers to write column headers with no spaces between words. In this activity, you'll see "DayOfYear" instead of "Day Of Year", and you'll write "Year_bins" instead of "Year bins". This is a standard convention because spaces between words can create problems when exporting the data set into some statistical analysis programs (which you won't do in this activity).

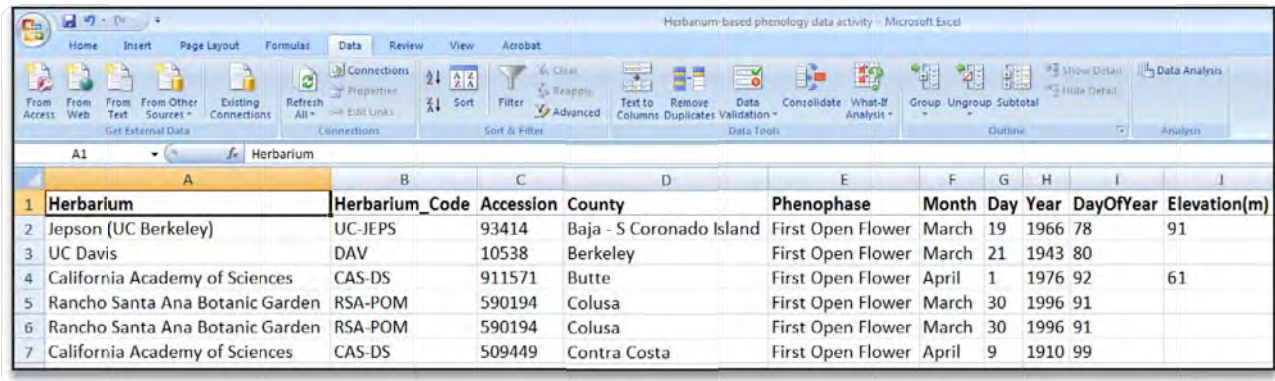
We also have included two columns to identify the herbarium from which each specimen was sampled. One column provides the full name of the herbarium, and the other column reports an abbreviation of the herbarium name. The abbreviations are routinely used by experienced botanists as a shorthand, but because this data set may be shared among students and instructors who are unfamiliar with them, it's important to include the full name as well. While many researchers use abbreviations to describe attributes of their data in spreadsheets, this can be a risky practice because the essential information can be lost as time passes and memories fade. Moreover, when data are shared with the public, it's helpful to minimize the amount of explanatory information that must be provided, and reducing the use of abbreviations within data sets can help to achieve this.

Outline of the procedure you'll follow to explore historical flowering phenology of California poppy

Page	Process	Procedures	Goal / Outcome	
7	Step 1	Download data set	Visit the USA National Phenology Network website	Open the file
7	Step 2	Organize data	Sort & Filter	Clean up the data set for next steps
9	Step 3	Summarize data	PivotTable & PivotChart	Determine the number of herbaria, the number of specimens per herbarium, and the number of collections per county in California
11	Step 4	Visualize data & format for presenting	Histogram	Create a frequency distribution of collection year, day of year, and elevation
19	Step 5	Examine data for long-term trends	Scatter plot Regression analysis	Create a scatter plot and add a trendline Conduct a regression analysis

STEP 1: DOWNLOAD THE DATA

Visit the Education section of the California Phenology Project (www.usanpn.org/cpp/education) or of the USA National Phenology Network (www.usanpn.org/education). Browse the available files for a Microsoft Excel file entitled “Herbarium-based phenology data activity”. Then download the data set, save it to your computer, and open the file. It should look like this once opened:



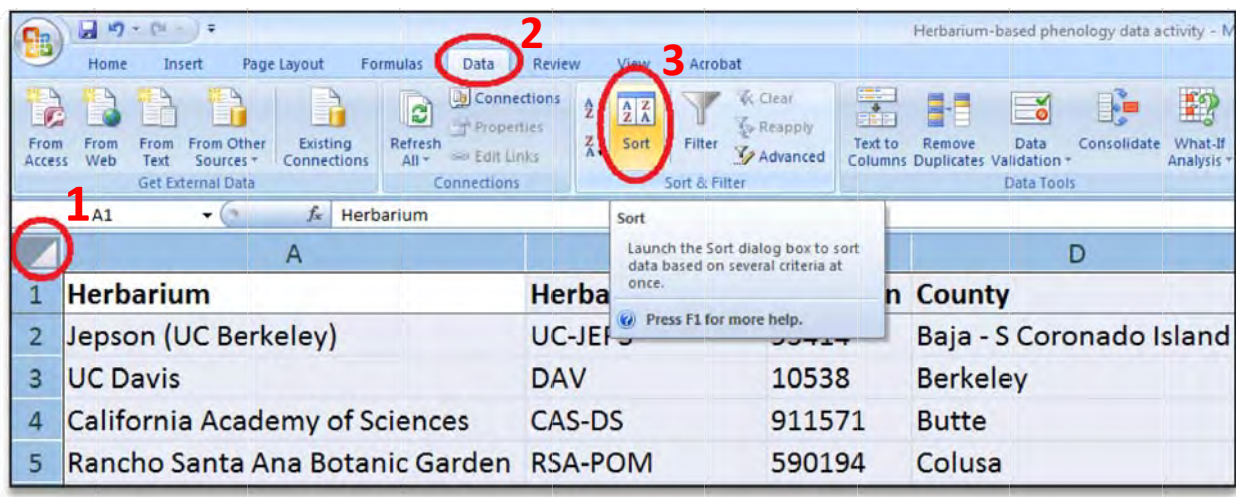
The screenshot shows the Microsoft Excel interface with the following data:

Herbarium	Herbarium_Code	Accession	County	Phenophase	Month	Day	Year	DayOfYear	Elevation(m)
Jepson (UC Berkeley)	UC-JEPS	93414	Baja - S Coronado Island	First Open Flower	March	19	1966	78	91
UC Davis	DAV	10538	Berkeley	First Open Flower	March	21	1943	80	
California Academy of Sciences	CAS-DS	911571	Butte	First Open Flower	April	1	1976	92	61
Rancho Santa Ana Botanic Garden	RSA-POM	590194	Colusa	First Open Flower	March	30	1996	91	
Rancho Santa Ana Botanic Garden	RSA-POM	590194	Colusa	First Open Flower	March	30	1996	91	
California Academy of Sciences	CAS-DS	509449	Contra Costa	First Open Flower	April	9	1910	99	

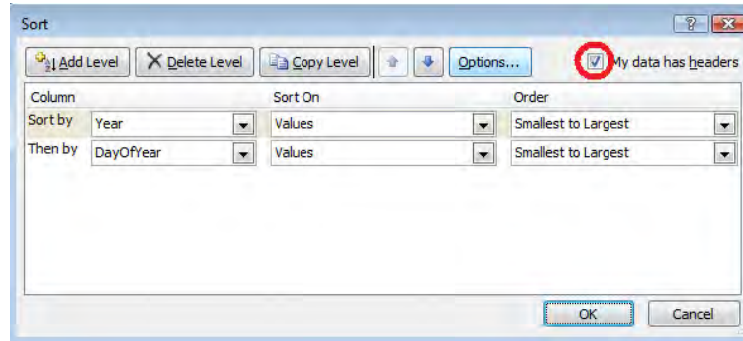
STEP 2: SORT AND FILTER THE DATA

Now you’re ready to clean up the data by sorting and filtering. Sorting data simply rearranges rows on the spreadsheet; filtering data actually hides unwanted rows. Both processes are useful to learn, particularly for the purposes of preparing a data set for analysis.

- Sort the dataset.** First highlight the entire data set by clicking on the upper-left corner of the spreadsheet (1). Next, select [Data > Sort]. The Sort window will appear (see next page for an image) – make sure the “My data has headers” box is checked, then sort by Year (on values, order smallest to largest). Then “Add Level”, and then sort by DayOfYear (on values, order smallest to largest). Click OK. All of the rows in the data set will become rearranged.



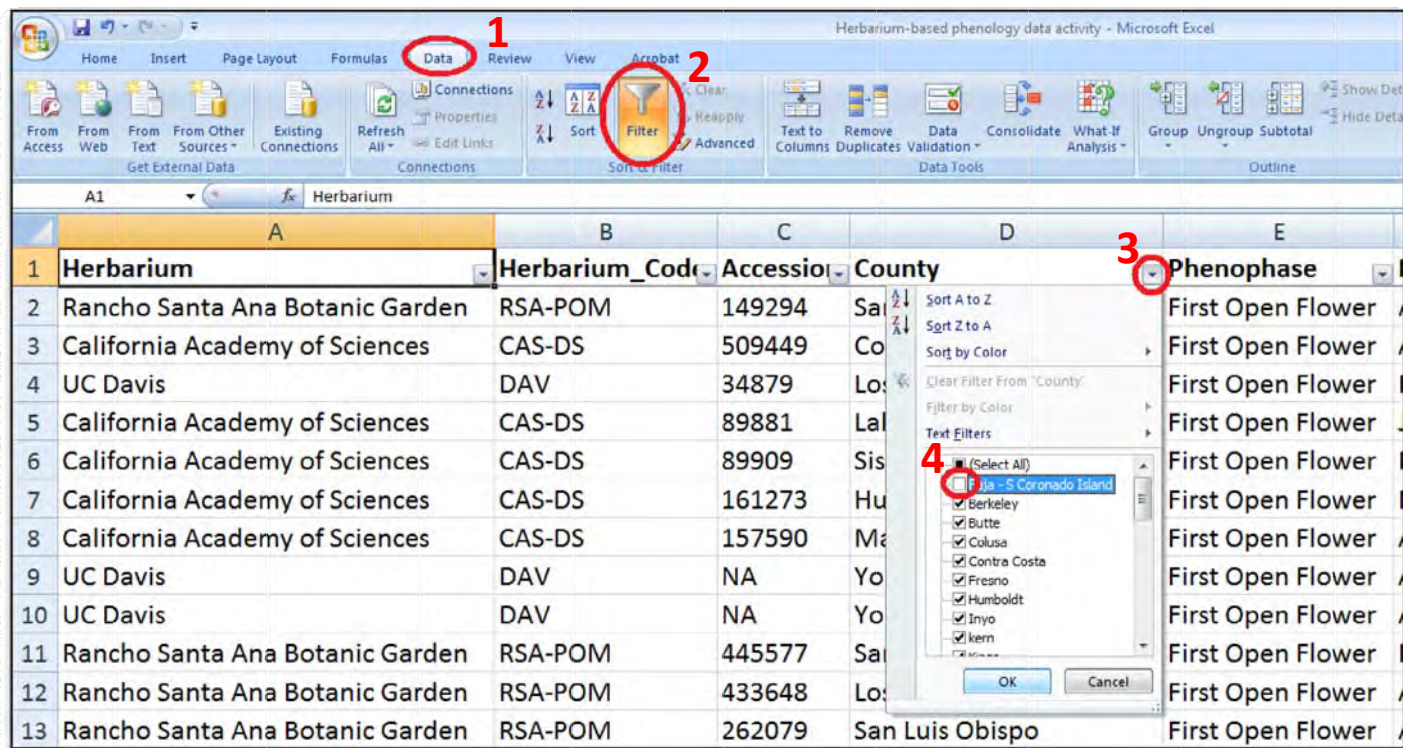
↓ Sort window shown on next page



Your data set should appear as below – check that Accession 149294 from RSA-POM is in cell C2, and that there are 149 rows of data (therefore, sample size N=148 because of the header row). **It is imperative to remember to sort the entire data set every time you sort – if you sort just a few columns independently of the entire data set then you’ll disassociate the data!** For example, if you sort just the DayOfYear column independent of the spreadsheet, then Accession 149294 in row 2 will be re-assigned to a DayOfYear of 64 (go ahead and try it – Excel should try to stop you from doing this, but go ahead to see what happens. Afterward, just hit CTRL+Z once to step backward, making sure your data set matches the image below before continuing).

	A	B	C	D	E	F	G	H	I
1	Herbarium	Herbarium_Code	Accession	County	Phenophase	Month	Day	Year	DayOfYear
2	Rancho Santa Ana Botanic Garden	RSA-POM	149294	San Francisco	First Open Flower	April	15	1906	105
3	California Academy of Sciences	CAS-DS	509449	Contra Costa	First Open Flower	April	9	1910	99
4	UC Davis	DAV	34879	Los Angeles	First Open Flower	May	1	1912	122
5	California Academy of Sciences	CAS-DS	89881	Lake	First Open Flower	June	8	1919	159
6	California Academy of Sciences	CAS-DS	89909	Siskiyou	First Open Flower	May	29	1923	149

2. **Filter out unwanted data.** Now that your data set is sorted by Year and DayOfYear, filter out unwanted rows of data. There are only 9 specimens representing the Peak Flowering phenophase (not enough power for a robust study), and there is one specimen from Baja (outside the geographic region of interest for this study). Instead of searching for these rows individually and deleting them, filtering provides an automated solution (see next page for image):
 - a. With any cell highlighted, select [Data > Filter]. A new symbol will appear in the corner of each header cell. Click on the symbol for the County column, un-check the box for Baja, then click OK.
 - b. Repeat these steps for the Phenophase column (un-check Peak Flowering).



Important note #1: The filtered data don't actually get deleted – they only become hidden until you clear the filter. Scan down to the bottom of the data set – there still should be 149 rows, however you should be missing several rows throughout the data set. The data are hidden until you select [Data > Filter] again.

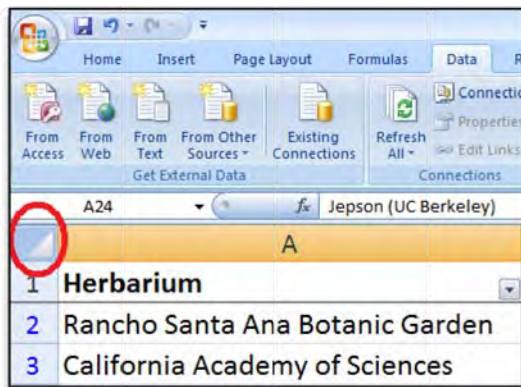
Can you bring back the filtered data? Yes. Just release the filter by clicking on [Data > Filter] again.

How do you know whether a column has been filtered? The symbol in the header row changes, and the row numbers below the header row are colored. Notice how the symbol for the County and Phenophase columns appears different than in the other columns. Click on the symbol and it will remind you how the data have been filtered. You also can adjust the filter again.

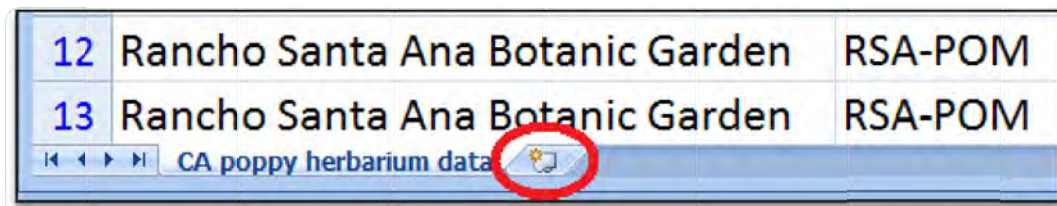
1	Herbarium	Herbarium_Code	Accession	County	Phenophase	Month
2	Rancho Santa Ana Botanic Garden	RSA-POM	149294	San Francisco	First Open Flower	April
3	California Academy of Sciences	CAS-DS	509449	Contra Costa	First Open Flower	April

Important note #2: The data visualizations and analyses that you will conduct in the next steps of this exercise do not necessarily recognize the data filter that you've just applied. The easiest way around this problem is to copy and paste the entire data set to a new worksheet. To do this:

1. Highlight and copy the data set either by:
 - a. Hold a left-click while highlighting the data set from cell A1 to cell M149, then either use [Ctrl+C] or right-click and select copy; or
 - b. Right-click the upper-left corner of the spreadsheet (circled below) and select copy.



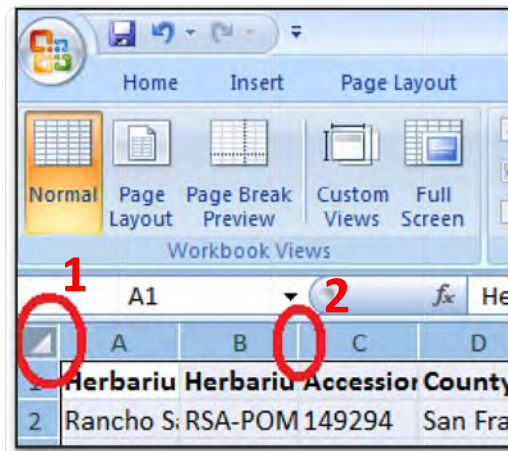
- Next, click on the 'Insert worksheet' symbol that is located at the bottom of the window, next to the title of your current worksheet 'CA poppy herbarium data'.



- A new worksheet should appear titled 'Sheet 1'. Right-click on cell A1 (the upper-left corner of the spreadsheet) and select Paste, or select cell A1 and use [Ctrl+V]. Your filtered data set is now pasted, excluding the rows that you filtered on the previous spreadsheet. You can check this by scanning down to view the total number of rows (it should be 139 rows including the headers). Right-click the title of 'Sheet 1' and rename this worksheet 'Filtered data for analyses' (see below).

136	UC Davis	DAV	154999	Yolo	First Open	March	16	2001	75
137	UC Davis	DAV	154999	Yolo	First Open	March	16	2001	75
138	UC Davis	DAV	92209	Yuba	First Open	April	14	1974	104
139	UC Davis	DAV	92209	Yuba	First Open	April	14	1974	104
140									
141									
142									
143									

- Notice how the column width is too narrow for some columns? There is an easy fix for this. Highlight the entire data set again (left-click the upper-left corner of the spreadsheet), and then double-click the dividing line between any two columns.



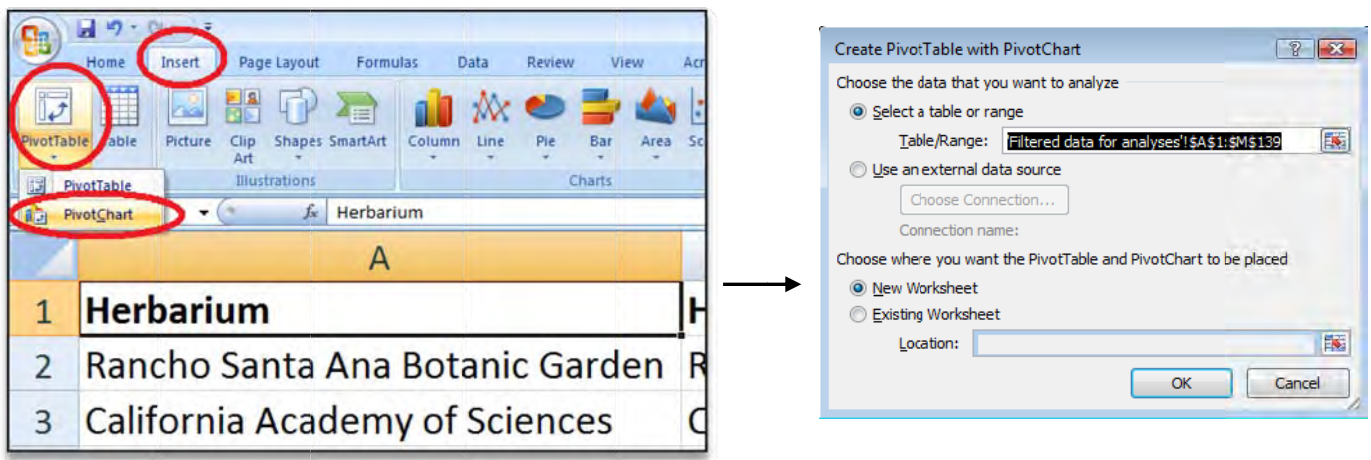
With your column width adjusted, you should now be ready to summarize, visualize, and analyze the data set.

STEP 3: SUMMARIZE THE DATA SET

Basic “summary information” about a data set is a standard part of any research project and presentation. How many herbaria are represented in this data set? How many specimens per herbarium were included? We’ll answer these questions not by counting row-by-row, but by using automated tools called PivotTables and PivotCharts.

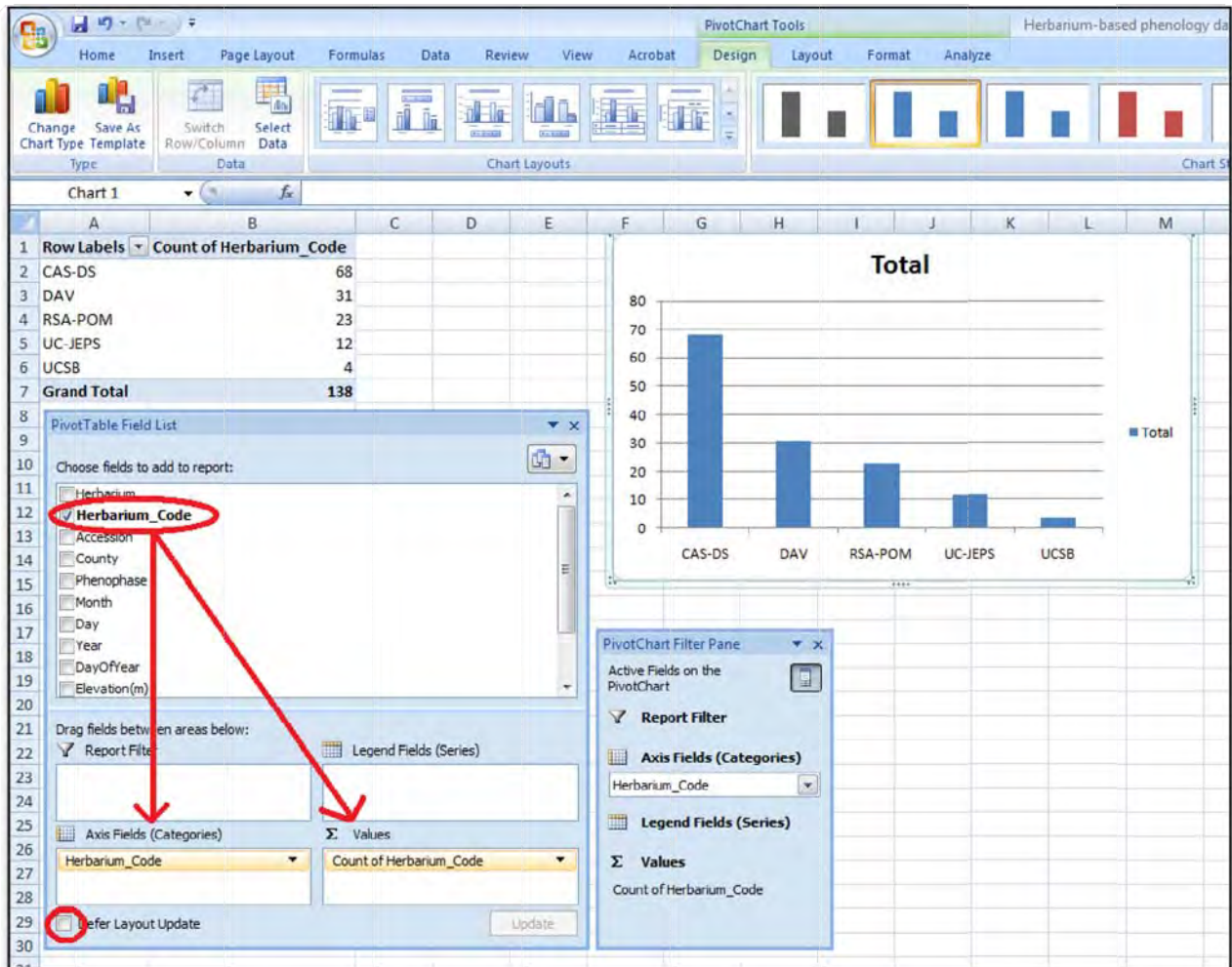
CREATE A PIVOTTABLE AND A PIVOTCHART

1. With any cell highlighted, select [Insert > PivotTable > PivotChart]. The “Create PivotTable with PivotChart” dialog box will open. By default, the entire data set will be selected (don’t worry about the \$ symbol in the range of data selected, this is fine). Make sure that “New Worksheet” is selected, and click OK.



- A new worksheet will appear – it looks a bit complicated, but don't be intimidated! Provide a new name for the worksheet, such as "PivotCharts". Arrange the pop-up windows so you can see each of them (see below for example), and then focus on the "PivotTable Field List" window. Be sure that the box in the bottom-left corner called "Defer layout update" is un-checked.

Check the box at "Herbarium_Code", and then grab the Herbarium_Code name and drag it to the "Axis Fields (Categories)" and the "Values" windows. The table should automatically fill in, and so should the chart. These are your PivotTable and PivotChart. Notice that the Grand Total is 138, indicating that your filtered data have, in fact, been left behind in your new 'Filtered data for analyses' worksheet. Play with the other categories of data as follows in order to explore the PivotChart function.



Select and drag Month into the "Axis Fields" window with Herbarium. **Determine the most frequently-represented month of collections in each herbarium.**

With Month and Herbarium in the "Axis Fields" window, change their positions so Month is above Herbarium. **How does this change the table and chart?**

What happens when you move herbarium to the "Legend Fields (Series)" window?

Now un-check Herbarium and Month, and instead check County. **Which county is represented most in the data set? Where is that county located?**

You should have noticed that “Axis Fields” and “Legend Field” are two different ways to view the same information. There is no right or wrong way to display the information, but the general rule is to keep tables and charts as simple and clean as possible. For our data set, Axis Fields seems to be the better option. In the next section of this activity you’ll have the opportunity to change the formatting on charts to improve not just their appearance but also the quality of the information conveyed. Leave your pivot chart set to summarize Herbarium and Month for now, and continue to the Step 4 to visualize the data in a histogram.

STEP 4: VISUALIZE THE DATA

One of the first steps all researchers take in analyzing data sets is visualizing the distribution of the data. It is important to understand the structure of the data before exploring trends in the data and interpreting results. Many statistical tests require that the data show a “bell-shaped” or “normal” distribution, with most data points clustered near the average (or mean) value of the distribution and the rest of the data points representing values that are more distant from the mean (including values that are both higher and lower than the mean). The values that deviate greatly from the mean are relatively infrequent, and appear as the “tails” of the bell-shaped distribution. For many statistical tests (e.g., regressions, t-tests, and the analysis of variance), if the data aren’t normally distributed, then the criteria for accepting or rejecting the hypothesis tested by the procedure will not be accurate. To visualize data you’ll make a histogram containing a *frequency distribution*, for several variables in the data set. Return to the data set and continue using filtered data from Step 1.

CREATE A HISTOGRAM FOR THE NUMBER OF SPECIMENS PER YEAR

- 1. Determine the smallest and largest values for the Year column – these will become the lower and upper limits on the x-axis of your histogram.** With the data sorted and filtered, scan through the Year column and identify the earliest and latest years in the data set. Alternatively, you could click on the filter pull-down menu for the Year column and scan that. Remember these two years (or write them down) before proceeding to the next step.

Tip: When scanning down the Year column to find the latest year, you’ll notice that the headers in row 1 disappear. If you want to keep the headers in view at all times, click [View > Freeze panes > Freeze top row].

- 2. Create bins for the term Year.** Bins are categories for the data – they will be the x-axis intervals in the histogram. For our histogram, each year will be a bin.

Establish a new column on the spreadsheet. Right-click the top of a column (we right-clicked column H containing DayOfYear). On the pop-up window, select Insert. A new blank column will appear in column H, and the Elevation column (and everything else to the right) will shift to the right to make room.

Type in the column header “Year_bins”, then adjust the column width to make room for the full header. In row 2, type in the earliest year in the data set (1906). Actually, for reasons you’ll see later

when the histogram is created, go ahead and type in a nice round number preceding your earliest data point... type in 1900. In row 3, type in the next year (1901), and continue filling in the Year_bins column until you reach the latest year in the data set (2009). Once again, go ahead and continue to the nice round number of 2010. To automate this fill-in process, just type in the first two years in the appropriate cells, highlight those two cells, then grab the bottom-right corner of the highlighted cells and drag the highlighted area down until the automatic counter reaches 2010. There will be blank cells below the 2010 cell – that’s correct.

Month	Day	Year	Year_bins	DayOf
April	15	1906	1900	105
April	9	1910	1901	99
May	1	1912		122
June	8	1919		159
May	29	1923		149

What happens if you highlight just one cell and drag it down?

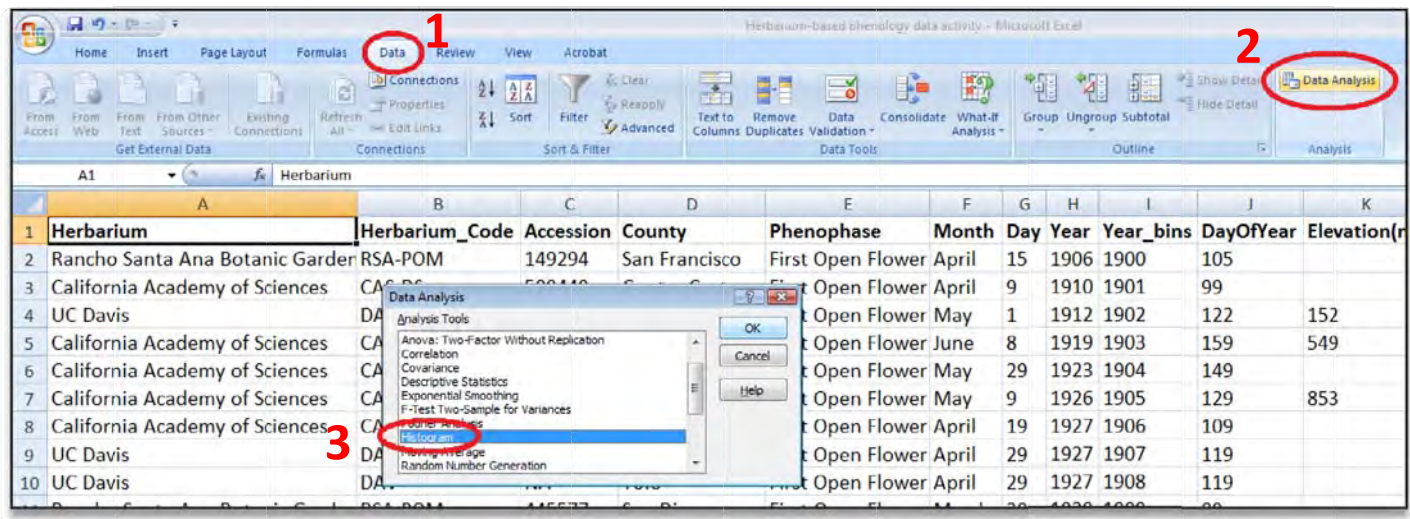
What happens if you highlight both cells but then grab and drag a different corner, or the side of the highlighted area?

Try double-clicking on the bottom-right corner of the highlighted cells. All the cells in the column should be filled in automatically. How far down does the automatic fill-in work? Did it stop at 2010 where you wanted? If it continued farther than 2010, when did it stop and why?

Now go back a step – clear the column, type in 1900 and 1901 again. This time, type in something a few cells down – anything. Then perform the same double-click automatic fill-in again... what happened?

Always be cautious with the automatic fill-in tools, and check their accuracy!

3. **Create the histogram.** With any cell on the spreadsheet selected, open the Data Analysis tool [Data > Data Analysis]. Select Histogram and click OK.

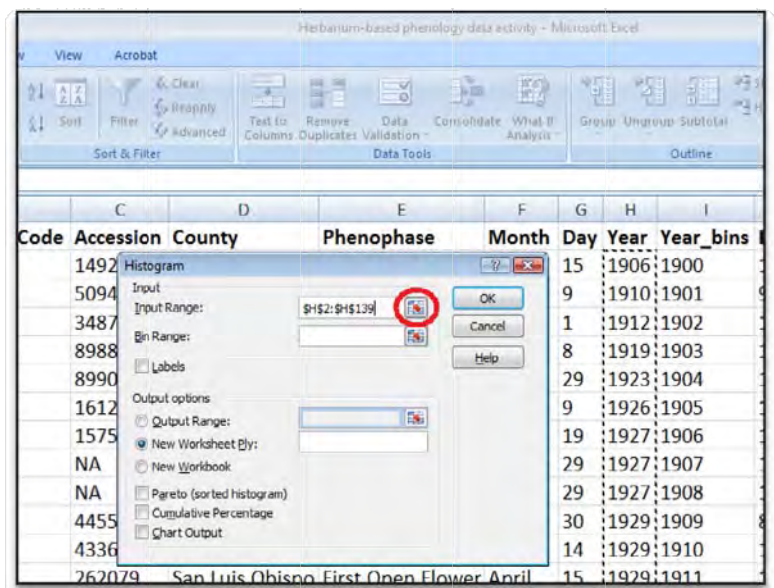


If you can't find the Data Analysis tool, don't fret! If you haven't used it before on your computer, you'll probably have to activate it. You only have to do it once; it takes just a moment following these steps: Click the **Microsoft Office symbol** at the upper-left corner of the window, and then click **Excel Options**. A new window should pop up. Find **Add-Ins** on the left column, click on it, and then click on **Analysis ToolPak** (you can ignore the Toolpak with the VBA option). Click **OK**. If you have trouble going through this process, or if you want to learn more about Add-Ins, consult the Microsoft Support website (search for Analysis ToolPak or Add-Ins).

- **With the Histogram dialog box open, select the Input Range.** Click on the Input Range symbol (circled in red below); then highlight all of the values in the Years column of the data set; and then click again on the symbol in the dialog box. The assigned input range should be from cell H2 to cell H139 (the \$ symbols will appear automatically – that's correct).

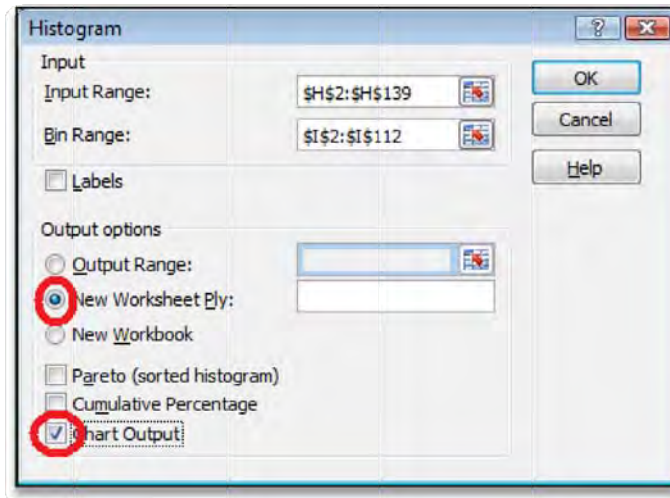
- **Select the Bin Range**

Repeat the same process for the Bin Range by clicking the "Bin Range" symbol in the dialog box; selecting the values in the Year_bins column of the data set (I2 to I112), and then clicking the symbol again to return to the Histogram window.



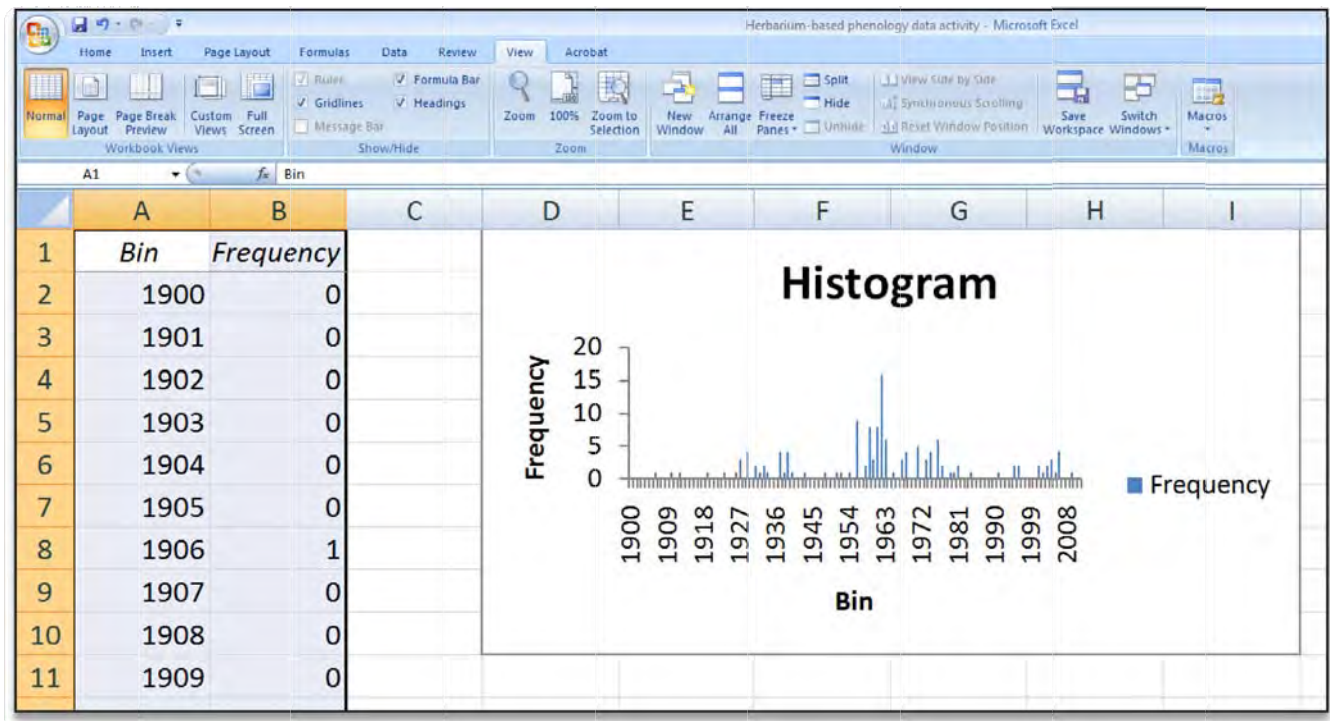
- **Select the Output Options**

Select “New Worksheet Ply:” and type in a brief title (for example, “Hgram by year”). Be sure to select “Chart Output”. Click OK.



- **View the histogram**

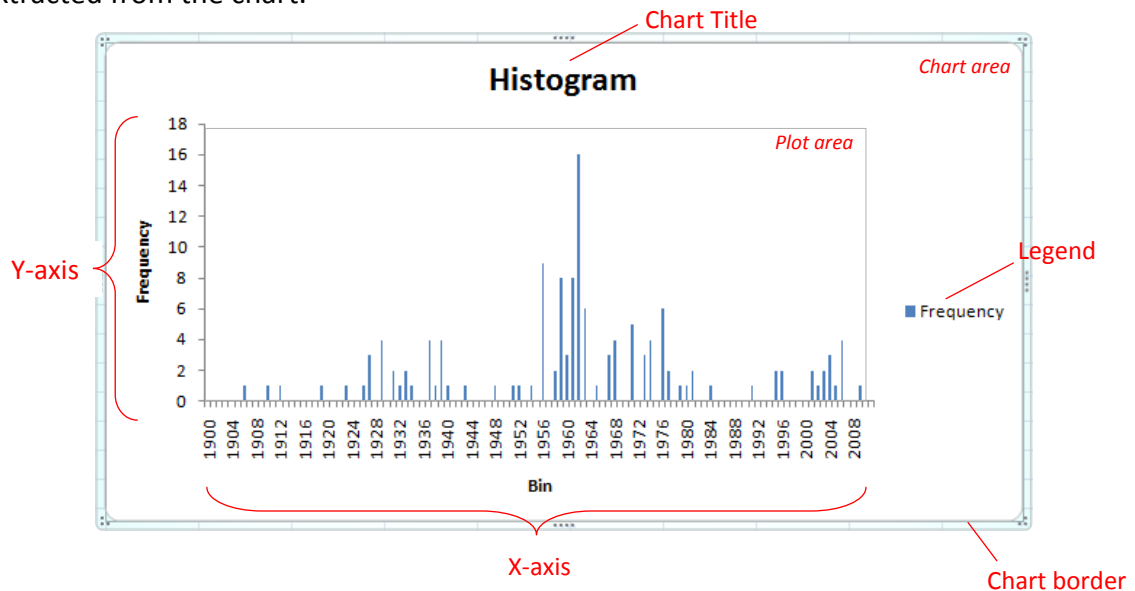
A new worksheet will appear with Bin and Frequency columns, and a histogram. *Voilà!* The histogram may be pretty small, so drag a corner to increase the size.



Tip: If you hold down Shift on your keyboard while dragging the corner, you’ll keep the same proportions as you increase size. This is generally true in all Microsoft applications and in many other graphic design programs.

FORMAT THE HISTOGRAM FOR INTERPRETATION, SHARING, AND PRESENTING

The histogram you produced is displayed in the default format. Though effective in visualizing the data, we should make a few brief formatting changes to make it appear cleaner while enhancing the information that can be extracted from the chart.



The area of the histogram with the vertical bars (bounded by a gray box) is called the “plot area”, and it essentially floats over the “chart area” (which includes the axes and their titles, the chart title, legend, background, and border). Pause your cursor over these areas to become familiar with their boundaries. Notice that you can left-click or right-click on each component of the histogram and format its appearance. Format the histogram as follows:

- **Titles** – Left-click the chart title and change it to “Frequency distribution of California poppy specimens in herbarium-based phenology activity”. Change the title font size to 16. Next change the x-axis title to “Year of specimen collection”. Change the axis title font size to 14.
- **Legend** – Left-click and delete the legend (while critical for other types of graphs, it is not necessary for this histogram).
- **Data series** – Right click any of the vertical data bars and select “Format data series...”. From the left column select “Fill”, then select “Solid fill”, then choose a dark bold color (we’ve selected a dark gray). Also format the “Border color” in the same way (we’ve selected black).
- **Axes** – Change several components of each axis by right-clicking on the axis and selecting “format axis...”. Do the following:

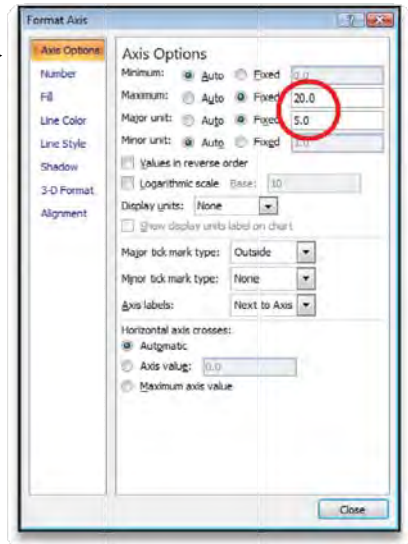
For the x-axis, under “axis options”:

- Change the “Interval between tick marks” to 5 years
- Change the “Interval between labels” to 10 years
- Briefly look at the other formatting options, but don’t get lost in the details. Remember – keep it simple and clean.



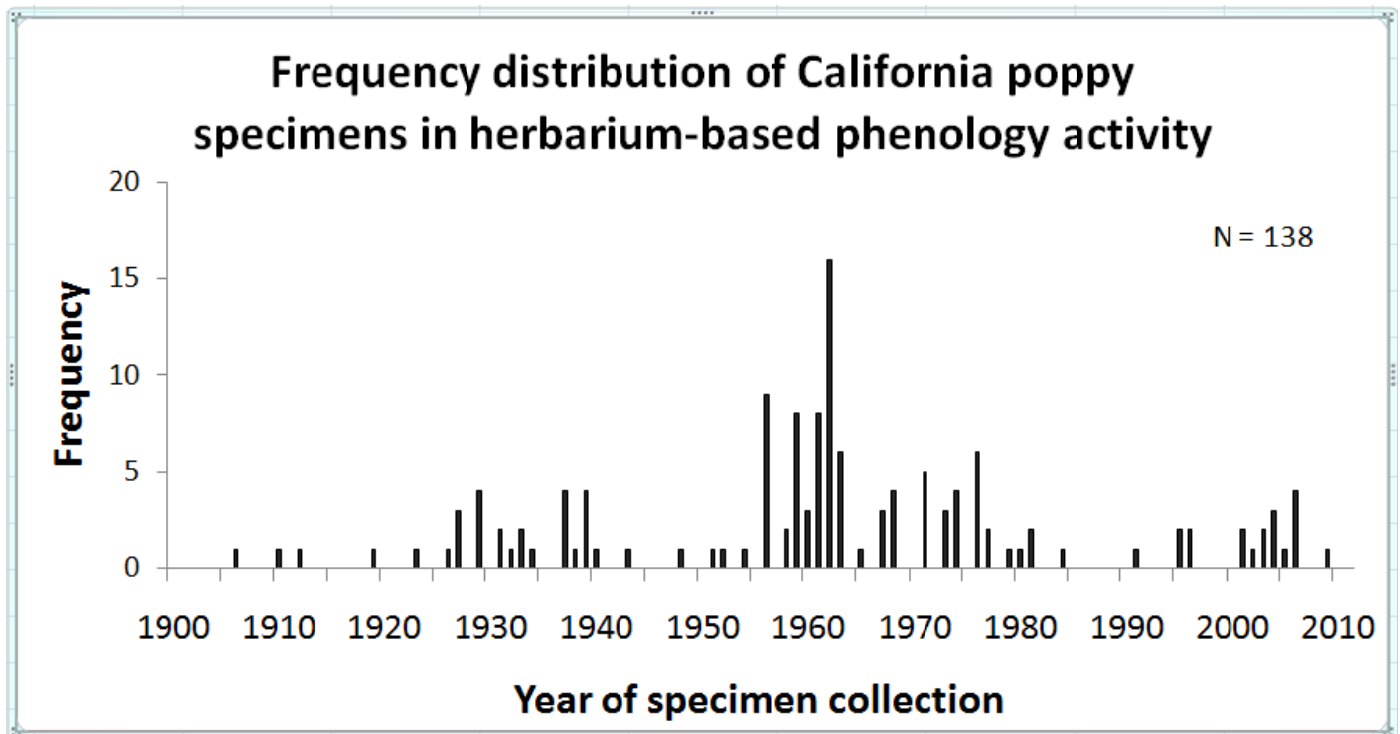
For the y-axis, under “axis options”:

- Change the “Maximum value” to fixed and type in 20
- Change the “Major unit” to fixed and type in 5
- Briefly look at the other formatting options, but don’t get lost in the details. Remember – keep it simple and clean.



Finally, making sure the chart area is highlighted (that is, click on the chart with the cursor rather than on a cell in the spreadsheet), select [Insert > Text Box]. Left-click and drag the cursor to create a small text box anywhere on the chart area. Type in “N = 138” (this is the sample size of the study after filtering). Drag the box to the side of the plot area.

Your chart should be formatted as follows:

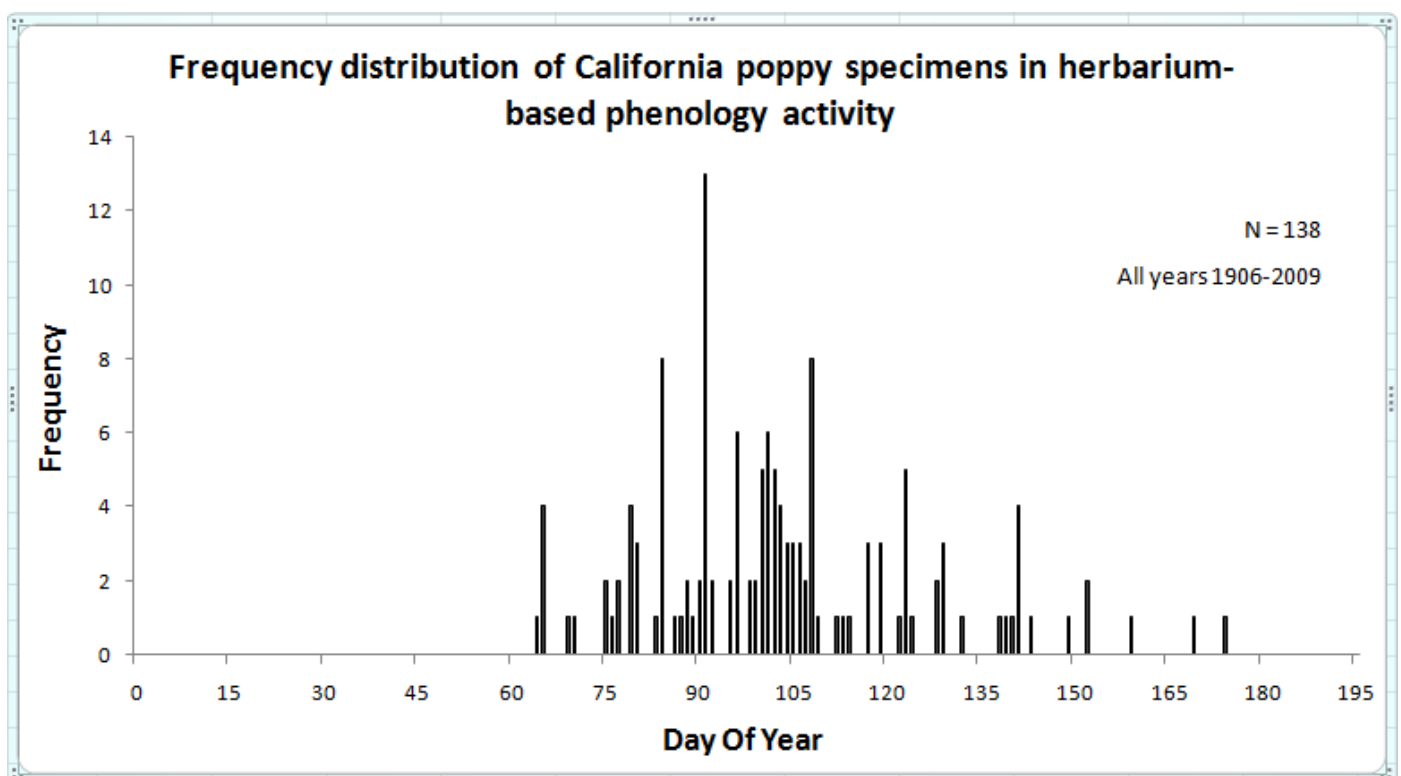


The histogram now appears cleaner and is easier to read because of simplifying changes such as longer axis intervals (but not too long), and fewer tick bars between intervals. In addition, the new titles and sample size help to describe the context for the data better than the default appearance.

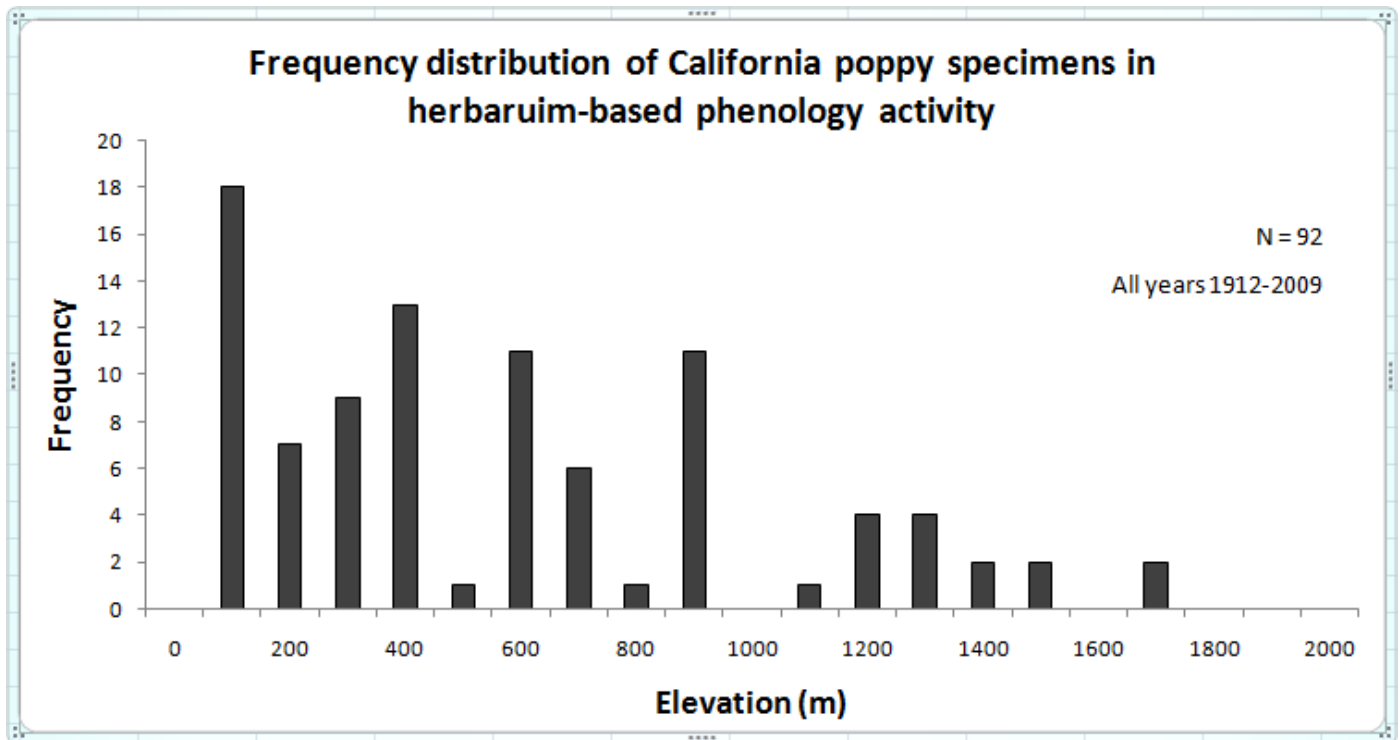
The distribution of the data indicates that the frequency (simply, the number) of specimens in the herbarium phenology data set varies by year and by decade. There is, more or less, good representation of specimens over the century, but some decades (1960's, 2000's) are better represented than others (1910's, 1940's, 1980's). *Do you think this might affect any trends we find in the next step? How?*

Now that you've completed the process of creating a histogram for one aspect of the herbarium phenology data set, you should be capable of creating other histograms from the same data set.

Following the same process as above, create a histogram for the DayOfYear variable. For this data visualization, you will ignore Year and just examine the seasonal distribution of California poppy collections. *When is the most frequent time of year for botanists to collect California poppy plants in the first flower phenophase?*



Now create a histogram for the Elevation variable. Sort the data set by Elevation, determine the lowest and highest elevation values, create an “Elevation_bins” column, and then create and format a histogram. Keep in mind that not all specimens in the data set have elevation values, so the sample size decreased to N=99 and the time span changed to 1912 – 2009.



Challenge: Interpret the distributions

How could the following factors affect the frequency distributions of California poppy specimens?

- The biology of California poppy
- Collectors
- Data collection method allowing for several individuals per sheet to be included
- Herbaria
- Historical socioeconomic events

How might the frequency distributions affect our interpretation of later results?

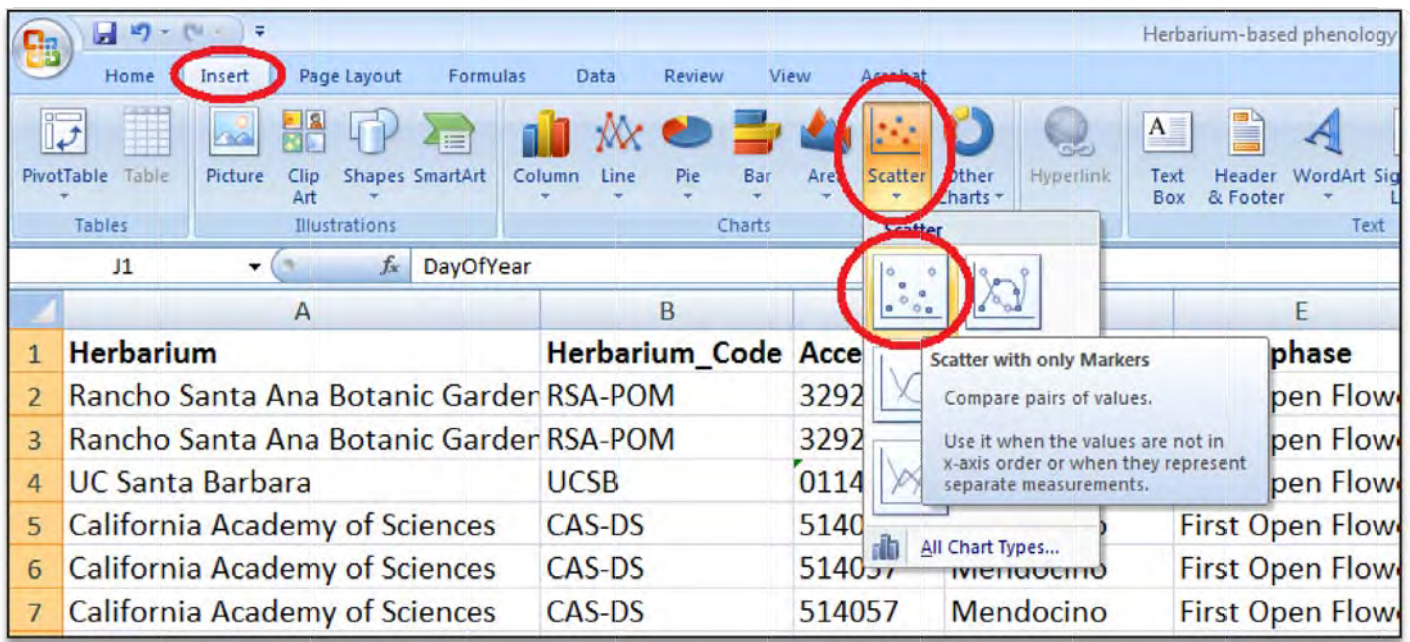
STEP 5: EXAMINE THE DATA FOR LONG-TERM TRENDS

Once the data set has been visualized and you have an understanding of the distribution of the data points, begin exploring the data for long-term changes in the flowering phenology of California poppy. Our filtered data set includes only plants that were collected in the same phenophase (first open flower), and long-term changes would manifest in temporal changes in the collection times. You will create a scatter

plot to visually inspect the long-term trend, and then conduct a regression analysis to assess the relationship between the DayOfYear and Year variables.

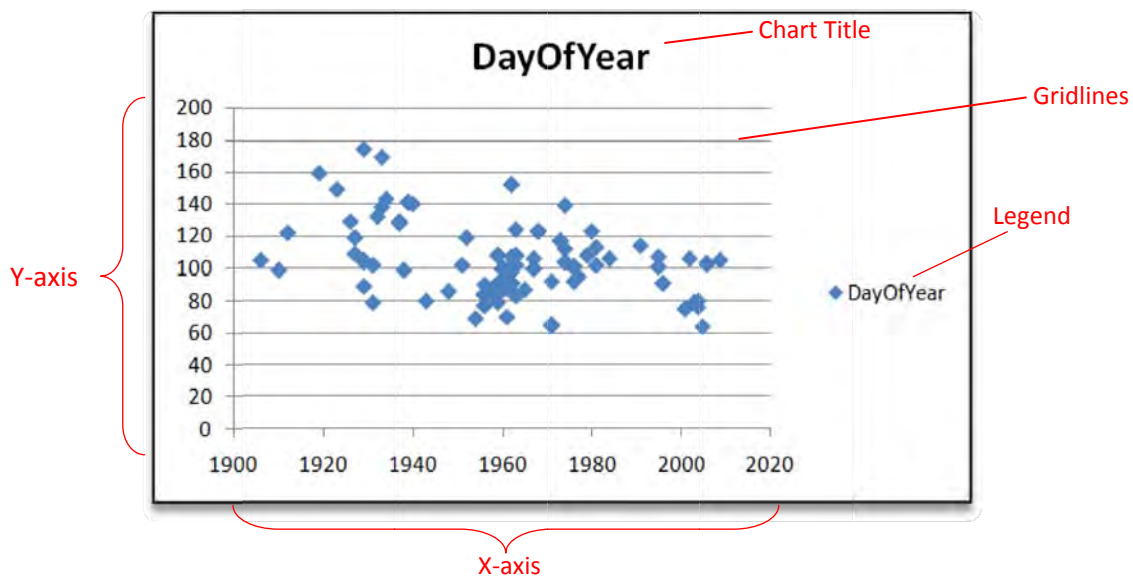
CREATE A SCATTERPLOT

Highlight the Year column (left-click column F). Then, while holding down the Ctrl key, highlight the DayOfYear column. Both columns should be highlighted. Next, select [Insert>Scatter> Scatter with only markers].



Notice that the Year column (not visible in this image) lies to the left of the DayOfYear column. This means that in the resulting graph, Year will be on the x-axis and DayOfYear will be on the y-axis. The standard convention in Excel is that the left highlighted column will be the independent variable (x-axis) and the right highlighted column will be the dependent variable (y-axis).

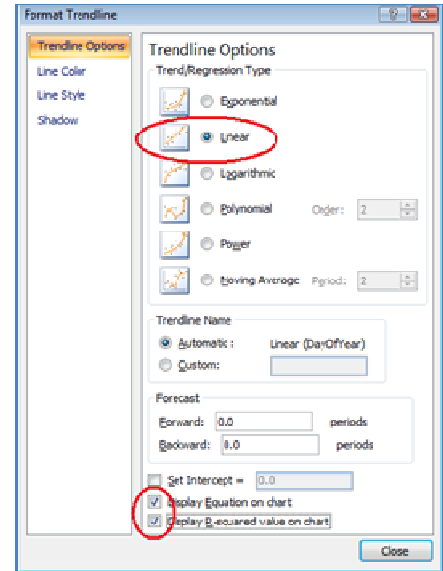
1. **View & format the scatter plot.** The graph should appear on the same worksheet and look like this:



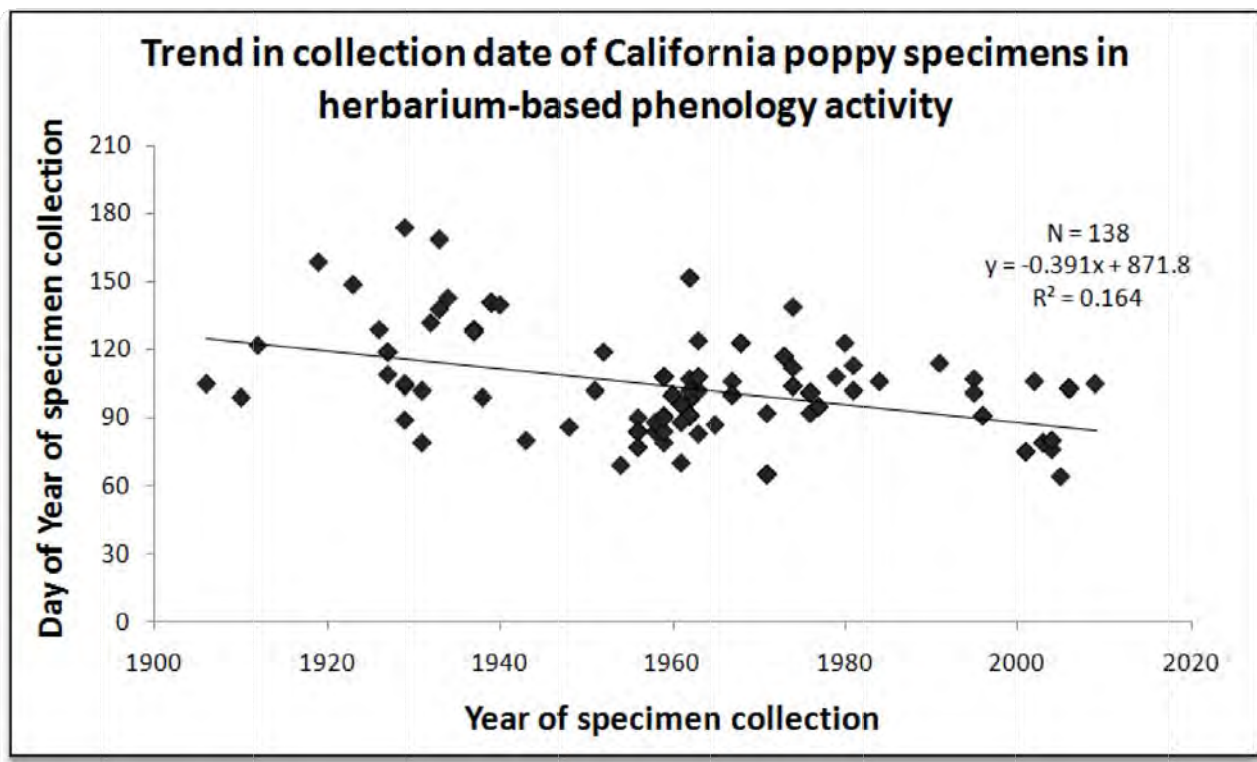
Formatting a scatter plot for clarity and simplification is similar to formatting a histogram – a few changes are all that’s needed to improve the clarity of the graph. Follow the same steps you took to format the histogram. Also, delete the gridlines by right-clicking and selecting delete. To add axis titles, select [Chart Tools > Layout > Axis Titles].

2. **Add a trendline to the scatter plot.** Right-click on any of the data points and select “Add trendline...”. In the left column of the Format Trendline window, select “Trendline Options” and then make sure that the following three components are selected:

- Trend/Regression Type – Linear
- “Display equation on chart”
- “Display R-square value on chart”



Finally, your scatter plot should appear similar to this:



A trendline also is commonly called a “line of best fit” because it is a line drawn through a cloud of data points that minimizes the sum of all of the distances (or “deviations”) between each data point and the trendline. The trendline is described by the general equation $y = mx + b$. In our data set, we already know x (Year) and y (DayOfYear), so we’re solving for m (the slope of the line). The value b is simply the y -intercept.

There are a few mathematical methods by which to calculate the distances from each data point to the trendline – the most common method minimizes the *vertical* distance between each data point and the trendline (called ordinary least squares). The R^2 value (spoken “R square”) represents the amount of scatter around this trendline – the more scattered the data, the closer to 0 the R^2 value becomes; the more linear the relationship (and the lower the scatter above and below the line of best fit) between the X- and Y-variables, the closer to 1 the R^2 value becomes. In other words, R^2 represents the amount of variation in the Y-variable (DayOfYear) that can be explained by variation in the X-variable (Year).

Challenge: Interpret the California poppy herbarium collection record trend

Include the trendline slope and the R^2 value in your answer. The most complete interpretations also will refer to the histograms previously created, and any other aspects of the data set that seem appropriate to discuss (and, potentially, to visualize). How could you incorporate factors from the previous challenge into your analysis and interpretation?

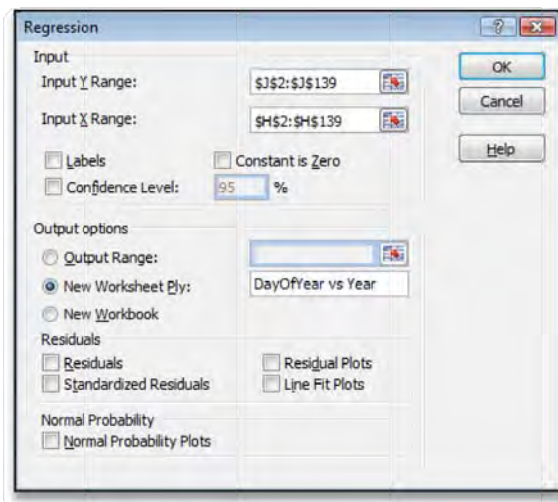
Example interpretation of the data – The data indicate a long-term trend toward earlier collection dates of California poppy, suggesting earlier flowering in recent years across California. Specifically, the slope of the trendline (-0.391) indicates that for every year (x) that passes, the California poppy plants included in this study and in the “first open flower” phenophase are, on average, collected about 0.4 days earlier, equivalent to about 4 days earlier per decade. This rate of phenological change is consistent with other reported values for observed wildflower phenological changes. In addition, the R^2 value indicates that about 16% of the variation in collection date can be explained by the year in which the plant was collected. Though not very high (recall R^2 ranges from 0-1), this R^2 value falls within a range commonly observed in ecological studies.

Note that, in this analysis, we did not control for possible effects of elevation or latitude on the phenological behavior of California poppy. We might expect, for example, that – all else being equal – plants collected from lower latitudes and elevations in California would flower earlier than those collected at higher latitudes and elevations. If this is the case, **and** if (by chance) more recently sampled plants happened to be collected from lower elevations or lower latitudes than plants sampled in the past, then this nonrandom distribution of elevations and latitudes across time could contribute to the temporal trend observed. How might this be controlled for in your data set?

CONDUCT A REGRESSION ANALYSIS

The ultimate goal of regression analysis is to determine our confidence level in stating that the slope of the trendline is not zero (sounds exciting, doesn't it?!). In other words, we need to take the important step of determining how confident we are that flowering phenology of California poppy is really changing.

1. With any cell selected, select [Data > Data Analysis] and then select "Regression" in the pop-up window.
2. Select input ranges and output options. Select the appropriate Y (DayOfYear) and X (Year) values, and provide a title for the New Worksheet Ply (regressions are commonly referred to as "Y versus X", so we used "DayOfYear vs Year").



3. View and interpret the regression Summary Output.

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.405361101							
R Square	0.164317622							
Adjusted R Square	0.158172899							
Standard Error	19.86128456							
Observations	138							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	10548.64003	10548.64003	26.7412562	8.13903E-07			
Residual	136	53648.0049	394.4706243					
Total	137	64196.64493						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	871.8428927	148.6778012	5.863974887	3.2609E-08	577.8235089	1165.862276	577.8235089	1165.862276
X Variable 1	-0.39177306	0.075760645	-5.17119485	8.139E-07	-0.541594335	-0.24195178	-0.54159433	-0.24195178

R² value is slightly lower than the previous calculation because regression is more conservative (and generally more accurate).

The "F-test" result indicates that there is a 0.0000023 chance that the regression line really does have a zero slope. In other words, we are precisely 99.999998% confident that this data set indicates California poppy plants are flowering earlier.

The slope of the regression line is reported as the coefficient of the x variable.

Challenge

Using the internet and/or any available statistical texts, learn more about the ANOVA (analysis of variance) and F-test that were a part of the regression analysis. Provide an interpretation of the regression results including a description of the reported standard error (located below the 'R Square' value in the *Regression Statistics* table) and lower/upper 95% confidence intervals (located near the bottom-right corner of the table).