

Tips and Guidance for Analyzing Data

Executive Summary

This document has information and suggestions about three things: 1) how to quickly do a preliminary analysis of time-series data; 2) key things to watch out for and compensate for when doing that analysis; and 3) the mechanics for doing some of these analyses in Google Sheets or Excel.

The document is divided into small sections and a table of contents is provided so that you can find material when it is needed. **One idea for getting started quickly** is to have some group member(s) start reviewing the “Tips”, “Notes”, and “Analyzing Data” sections while others focus on getting data from your sources and on the “Mechanics” sections.

Table of Contents

Table of Contents

| | |
|---|-----------|
| Top Ten Tips for Analyzing Data When Time is Limited..... | 2 |
| Note – Use Scatter Plots, not Line Graphs..... | 3 |
| Note – Google Sheets and Excel | 3 |
| Mechanics – Logging in to a Chromebook | 3 |
| Mechanics – Getting data from the Project Page into Excel or Google Sheets..... | 3 |
| Mechanics – Getting data from the Audubon Site into Google Sheets | 4 |
| Analyzing Data – Aligning Your Independent Variable ... Time | 4 |
| Analyzing Data – Taking Care with Dependent Variable Values (3 Situations)..... | 5 |
| Analyzing Data – Analyzing a Trendline (Least Squares Regression)..... | 6 |
| Analyzing Data – Calculating a “Rolling Average” (a.k.a. “Running Mean”) | 6 |
| Analyzing Data – Calculating a “Rolling Median” | 7 |
| Mechanics – How to Get a Scatter Plot Quickly and Easily..... | 8 |
| Mechanics – Pro Tip for Getting Multiple Graphs in Minimal Time | 9 |
| Mechanics – How to Add a Linear Trendline, Equation and R-Squared Value | 9 |
| Mechanics – How to Experiment with Non-Linear Trends & Tweak Format | 10 |
| Mechanics – How to Add Error Bars or Data Labels | 10 |
| Mechanics – How to graph a Rolling Mean or Median on the Same Graph as a Scatter Plot | 11 |

Tips and Guidance for Analyzing Data

Top Ten Tips for Analyzing Data When Time is Limited

1. Get your data quickly into a spreadsheet in columns.
 - Each sample should be a row.
 - Each measurement (or calculation based on the measurements) should be a column.
2. Add extra columns to quickly calculate values required to align different data with a common independent variable.
 - Often the independent variable is time, but it can be sample numbers or other things.
 - Aligning sample may mean choosing a time or date to be T=0 (or Year 1...).
3. Add extra columns to calculate any modifications of the data you may want to analyze
 - Examples include: rolling averages (a.k.a. running mean), rolling median, differences from the mean or median, or differences from another baseline measure
 - These are often very helpful. Try them. The formulas are easy to create.
4. **Graph your data using scatterplots, where each dependent variable column should be a series.**
5. Add regression trendlines to your data and display the resulting trendline equations and r^2 -values.
 - Unless you have a reason to believe that the underlying effects are non-linear, it is common to start with a linear regression.
6. Examine your data and results for preliminary conclusions and ideas for alternate analyses.
7. Only if appropriate, examine other regression patterns (e.g. linear, exponential, logarithmic, polynomial, ...). **Note that this is rare.** It is ONLY appropriate if you have reason to believe that the data should fit another pattern and can articulate that reasoning (e.g. amoeba population growth is expected to be exponential).
8. Iterate on steps 2-7 as many times as needed.
9. Once interpretations or conclusions have been drawn, “clean up” tables or graphs for presentation or display.
 - Focus on readability and clarity without compromising integrity or accuracy.
10. Finalize conclusions and key presentation points.

Be Smart!

When you log in to google or other personal accounts on a public device, TURN OFF (uncheck) the “Stay Logged In” button.

Note – Use Scatter Plots, not Line Graphs

For the type of analysis required in this project, you will want to use scatter plots, not line graphs. This allows the calculation of regression lines and associated information and it avoid meaningless “connect the dots” representations of your data.

Note – Google Sheets and Excel

The specific instructions in this packet mostly use Google Sheets for the benefit of students working on Chromebooks or other devices without access to MS Office. That said, MS Excel will do everything described in this document (and more). Commands are similar and should be relatively straightforward to find. Use the Command Control Button (green plus sign in upper right) when trying to format and add elements to graphs in MS Excel.

Mechanics – Logging in to a Chromebook

1. Turn on the machine using the power button inside on the upper right.
2. Choose the “browse as guest” option in the lower left part of the screen.
 - a. ... or use your KAMSC login if appropriate
3. The system will open Chrome in guest mode and you are ready to go.

Be Smart!
When you log in to google or other personal accounts on a public device, TURN OFF (uncheck) the “Stay Logged In” button.

Mechanics – Getting data from the Project Page into Excel or Google Sheets

1. If using Google sheets, have someone in your group login to google and go to their google drive.
2. If you are working on a topic where the data has been provided as a spreadsheet on the “Big Data 2022” Project Page, such as Phenophases or Lake Ice, start by navigating to this URL:
<https://kamscmr.com/weebly.com/kamsc-big-data-2022.html>
3. Scroll down and find the spreadsheet you want.
4. Get the file using the appropriate method for your system:
 - In **Windows** or **Mac OS**, right click (ctrl click for Mac) the file, choose “save as” and download the file to an appropriate folder (choose your folder so that you’ll be able to find the file). If using google sheets, then go to google drive click the blue “NEW” button in the upper left, choose “file upload” and upload the file. You can then open the file in Google Sheets from your Google Drive.
 - On the **Chromebook**, double click the link and the file will open in Google Sheets. Make sure you are already logged in to Google Drive. Then click the blue “Share” button in the upper right corner and choose “save as google sheets”. You should then see a “Last edit was ...” indication in the center of the sheet near the top of the page.

Mechanics – Getting data from the Audubon Site into Google Sheets

The instructions for Audubon Bird Count data activities are provided in the “Audubon Bird Count Activity” file, which is linked on the Weebly webpage for Big Data Days. Please access these and follow the directions there.

National Data for the assigned species are provided in Excel files that are linked on the Weebly site. Download the files for your species and then open them in Excel to start your analysis. Data on additional species and Michigan-specific data for your assigned species are available on the Audubon site. Detailed instructions for accessing these are provided in the aforementioned “Audubon Bird Count Activity” document.

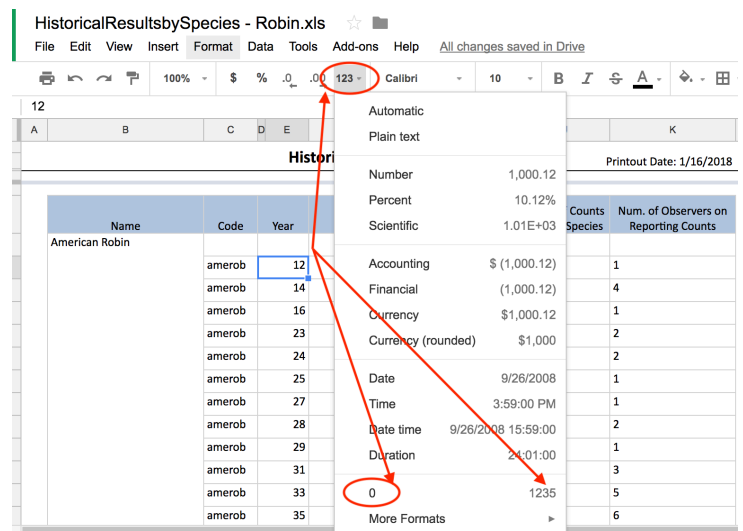
As described in those directions, the Audubon data is based on counts of sightings of specific bird species. Here are some keys to understanding this data:

- The fields are laid out as follows:
 - Year 1 is the year 1900. Year 115 is the year 2014.
 - The “Number” field is the raw number of birds that were counted that year.
 - The “Number/Party Hours” field is the data that is normalized for year-on-year comparison. It is the “Number” field divided by the number of hours the people doing the counting (the “Party”) spent observing.
 - The other data relates to the actual parties reporting.
- For several of the species, some of the early data (1st 10-40 years) is “noisy” or contains zeros (usually due to party hour data issues). In these situations, consider analyzing your data both with and without the “noisy” period to see what impact it has.
 - Also, remember to remove the zero years so that they don’t impact the analysis of your scatter plots (not collecting data is different than there actually being no birds).

Analyzing Data – Aligning Your Independent Variable ... Time

For the data sets in this project, your independent (horizontal axis) variable will normally be time. In many cases, you will be able to use this data unchanged. However, these are considerations you may want to take into account.

- Make sure your data for years, year numbers, or time periods are actually in “number” format. Avoid “general” or other formats that might cause the graphs to interpret your values as text.
 - Note: when Excel or Sheets shows numbers just starting at zero or one and counting on the horizontal axis of a scatter plot, it means that either no x-axis values were provided or that Sheets is interpreting the values provided as text instead of numbers.



| Name | Code | Year |
|----------------|--------|------|
| American Robin | amerob | 12 |
| | amerob | 14 |
| | amerob | 16 |
| | amerob | 23 |
| | amerob | 24 |
| | amerob | 25 |
| | amerob | 27 |
| | amerob | 28 |
| | amerob | 29 |
| | amerob | 31 |
| | amerob | 33 |
| | amerob | 35 |

HistoricalResultsbySpecies - Robin.xls

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

100% \$ % .0 .00 123 Calibri 10 B I U A . . .

Automatic
Plain text
Number 1,000.12
Percent 10.12%
Scientific 1.01E+03
Accounting \$ (1,000.12)
Financial (1,000.12)
Currency \$1,000.12
Currency (rounded) \$1,000
Date 9/26/2008
Time 3:59:00 PM
Date time 9/26/2008 15:59:00
Duration 24:01:00
0 1235
More Formats

Printout Date: 1/16/2018

| Counts Species | Num. of Observers on Reporting Counts |
|----------------|---------------------------------------|
| | 1 |
| | 4 |
| | 1 |
| | 2 |
| | 2 |
| | 1 |
| | 1 |
| | 2 |
| | 1 |
| | 3 |
| | 5 |
| | 6 |

- If you are analyzing more than one set of data, it may make sense to choose a year (or month) to be T=0 and use the same T=0 for all of your data sets. This can help you interpret things like y-intercepts on a consistent basis.

- For example, if you choose 1950 to be year zero, then 1960 would be year 10 and 2017 would be year 67.
- To do this, create a column for “Year #” and compute the values in it by taking your actual year and subtracting your zero year.

| | A | B | DA |
|---|------|--------|----|
| 1 | YEAR | YEAR # | |
| 2 | 1961 | 1 | |
| 3 | 1962 | 2 | |
| 4 | 1963 | 3 | |
| 5 | 1964 | 4 | |
| 6 | 1965 | 5 | |

- Pay attention to time gaps in your data. A normal scatter plot will simply exclude the missing data and can still produce a useful trendline. However, when computing a rolling average or rolling median, you need to take missing years into account. The easiest way to do this is to simply insert blank lines for missing years. Your scatter plots will ignore them and your rolling data will also if you are dragging an “average” or “median” formula down a column. Note that your averages and medians for some years will still be based on fewer data points, but they will at least span the correct number of years.

Analyzing Data – Taking Care with Dependent Variable Values (3 Situations)

Many of the data sets in today’s project already contain dependent variable values that are in correct, comparable units. Some may not, though, and it is important to make sure which are which. **Here are three situations in which your data series might require extra thought:**

1. If your data does not have the same or very similar number of observations in each major period over the time you’re examining, consider how you might adjust or at least understand what impact that might have on your results.
 - a. For example, if you have 12 data points per year for recent data and 1 point per year for older data, you might want to look at yearly averages.
 - b. Or, another example, if you have many missing data points in early periods, you have to be careful (as discussed in the previous section) when computing rolling averages.
2. If your data is not measured the same way all the time, you may need to be very careful with units and “normalize” the data.
 - a. In the Audubon data, for example, you will notice that the Audubon site graphs “Birds per Party Hour”. This is a great example of “normalized” data. The number of total hours observers spend observing a given species varies dramatically from time to time and from observation site to observation site. By calculating the average number of birds seen by each observer for each hour of observation, these effects are “normalized” out and the data becomes comparable.
3. If your data has significant outliers, it may be interesting to compare data points to a median or look at rolling medians, possibly comparing those to similar measures using the mean. The mechanics for doing these calculations are discussed in a later section of this document.
 - a. For example, the trendline of a data set and the trendline for the difference of those data and their mean (average) should produce very similar growth rates.

- b. If there are a lot of outliers, though, a trendline for the difference between the trendline for the data set and the trendline for the difference between those data and their median may suggest a different growth rate.
- c. Similarly, a rolling median may be less volatile than a rolling mean, again depending on the influence of outliers.

Analyzing Data – Analyzing a Trendline (Least Squares Regression)

An excellent starting point for looking at data over a time series is often to analyze a regression trendline. You've all done this many times, but here are a few reminders to keep in mind regarding trendlines.

- The trendlines we compute with Excel or Sheets (and in most situations) are “Least Squares Regressions”. We most often start with linear trendlines. They are so-called “best fit” lines where “best fit” is defined to mean that the sum of the squares of the y- (vertical) distances between the data points and the trendline is as small as possible.
- A linear trendline has an equation in $y = mx + b$ form, which gives you a slope and y-intercept that you can interpret in terms of your actual data and units.
- A regression trendline also has an associated r-squared (r^2) value, which in broad (approximate) terms measures the percentage of the variation in the data that is explained by independent (x-) value. It is one way to think about how strong the relationship in the data matters.
- Regression trend-curves can also be calculated using other underlying assumptions about the “shape” of the data. If you know or suspect that your underlying relationship may be explained by a polynomial, exponential or logarithmic function, for example, those trends can be calculated and graphed and their r^2 -values can be compared with a linear regression or each other. However, **you should NOT choose a shape JUST because the r^2 is better, there should be a reason you believe the underlying curve is appropriate.**
- The validity of a regression analysis depends on many factors that are beyond the scope of these notes. One key factor, though is the number of independent data points. In general, having more data points improves the meaningfulness of a regression-based trend. Things that increase the number of data points needed include: lower measurement precision, non-linearity of the underlying relationship, rapid changes in data (high volatility or steep slope), and others.
- These notes are NOT a substitute for a statistics course and a real understanding of regression. Rather, these are simply some practical observations that may be of use in your work today.

Analyzing Data – Calculating a “Rolling Average” (a.k.a. “Running Mean”)

Another often-helpful method of looking for trends in data is to “smooth out the data” (definitely laymen’s terms) by computing a rolling (or running) average (or mean), as in the following example:

- Suppose we have an annual measure of rainfall for the city of Grand Rapids and suppose those data vary dramatically from year to year.

- We could plot the data itself, but we could also plot several different rolling averages.
- For example, we could, starting with year 3 of our data, calculate a 3-year **trailing rolling average**. For each year, we would simply calculate the average of each year and the two prior years and then plot those data instead of the original measurements. In doing this we give up two data points, but we get a view of our data with reduced volatility.
- We can do the same thing for any period, 5 or 10 years, for example. The more data points we have, the more easily we can accommodate this. A 10-year rolling average on 50 years of data, for example, will be more useful than a 10-year rolling average on 20 years of data. We also give up n-1 data points for an n-year rolling average.
- We can also do a **central or forward rolling average** instead of a trailing analysis.

- In our rainfall example, a 5-year central rolling average would compute the average of this year with the data points from two prior years and the next two years. We would then plot that. In this example, we would give up two data points at the beginning of the series and two at the end.
- To get a future rolling average for the same data, we would average this year's data point with the next four (future) year's points and then plot those. In this example, we give up 4 data points at the end of the sequence.

fx **=average(D2:D4)**

| | A | B | C | D | E | F |
|---|------|--------|--------|-------|---------------------------|------------------------------|
| 1 | YEAR | YEAR # | DATE | DAY # | 3-YR CENTRAL ROLLING MEAN | 5-YR TRAILING ROLLING MEDIAN |
| 2 | 1961 | 1 | 6-May | 126 | | |
| 3 | 1962 | 2 | 3-May | 123 | =average(D2:D4) | |
| 4 | 1963 | 3 | 5-May | 125 | 124.33 | |
| 5 | 1964 | 4 | 4-May | 125 | 128.67 | |
| 6 | 1965 | 5 | 16-May | 136 | 135.67 | 125.00 |
| 7 | 1966 | 6 | 26-May | 146 | 144.00 | 125.00 |
| 8 | 1967 | 7 | 30-May | 150 | 143.00 | 136.00 |
| 9 | 1968 | 8 | 12-May | 133 | 139.33 | 136.00 |

- In all cases, once you have defined the formula (as shown in the figures), you can simply drag the formula down the length of the column to calculate the rolling data for the entire data set. To drag a formula, hover over the lower right corner of the cell you want to drag until you see the small solid square appear there, then click and drag that square.

Analyzing Data – Calculating a “Rolling Median”

As discussed previously, a rolling median is sometimes interesting to compare to other analyses because it can effectively mute the effect of outliers, sometimes “smoothing” data even more than a rolling mean.

- This is, however, a double-edged sword. It is valuable because it can provide insights to help begin to understand potential underlying trends in a data set, but it is also important to consider that outliers and volatility are part of the real data and may be important to a correct understanding of the real dynamics of a relationship.

fx **=MEDIAN(D2:D6)**

| | A | B | C | D | E | F | G |
|--|------|--------|--------|-------|-------------------|-----------------------|---|
| | YEAR | YEAR # | DATE | DAY # | 3-YR ROLLING MEAN | 5-YR ROLLING MEDIAN | |
| | 1961 | 1 | 6-May | 126 | | | |
| | 1962 | 2 | 3-May | 123 | | | |
| | 1963 | 3 | 5-May | 125 | 124.67 | | |
| | 1964 | 4 | 4-May | 125 | 124.33 | | |
| | 1965 | 5 | 16-May | 136 | 128.67 | =MEDIAN(D2:D6) | |
| | 1966 | 6 | 26-May | 146 | 135.67 | 125.00 | |
| | 1967 | 7 | 30-May | 150 | 144.00 | 136.00 | |
| | 1968 | 8 | 12-May | 133 | 143.00 | 136.00 | |
| | 1969 | 9 | 15-May | 135 | 139.33 | 136.00 | |

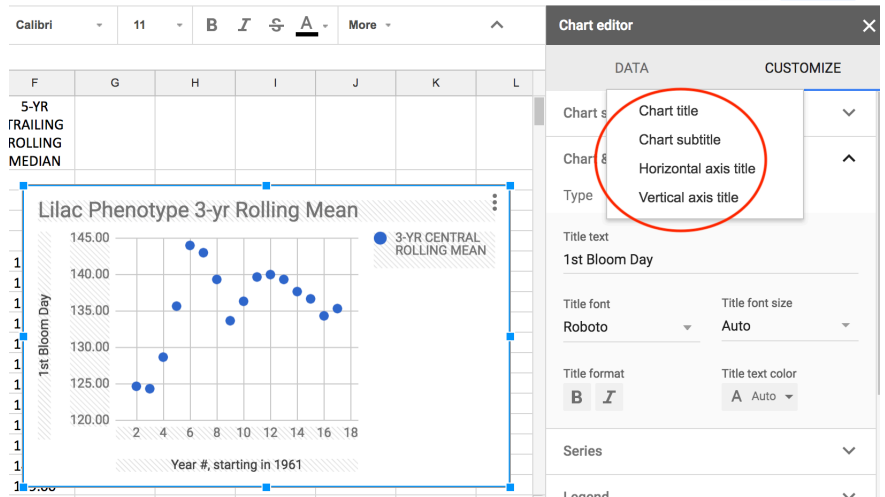
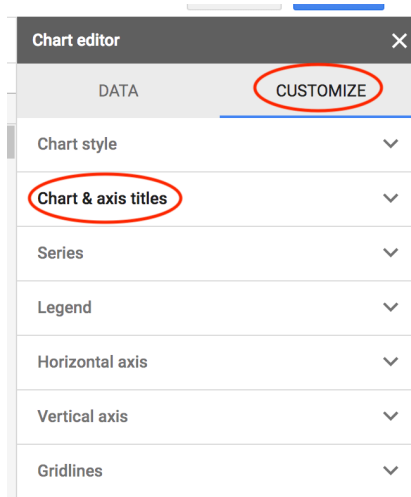
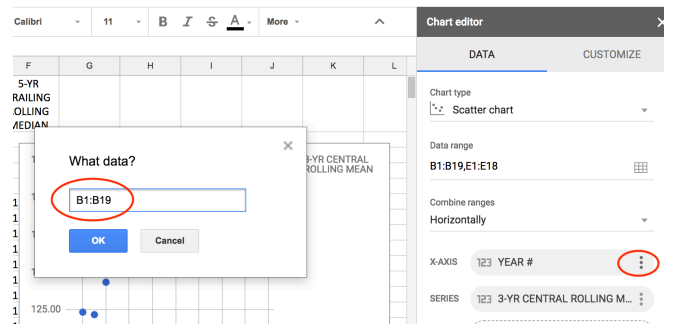
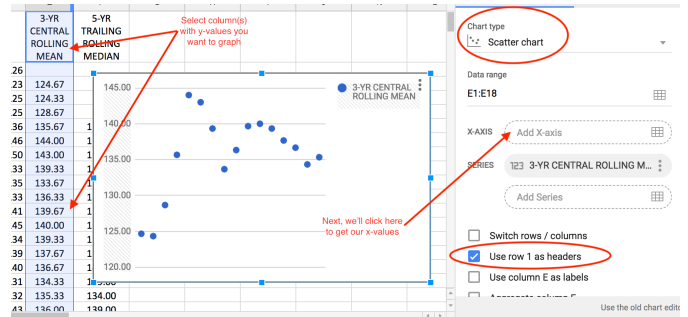
- Rolling Medians are calculated exactly like rolling averages, except using medians instead of means (obviously). They can be trailing, central, or forward, just as with rolling averages.

- In the long run, rolling means and medians are useful for gaining insight and sometimes very useful for communicating information to others, but they are not a substitute for more advanced statistical analyses.

Mechanics – How to Get a Scatter Plot Quickly and Easily

To easily create a scatter plot in Google Sheets, first organize your data so that your x-values are in one column (usually years, some other time measure, or a sample number) and your y-values (whatever you're measuring) are in another column. Then follow these steps:

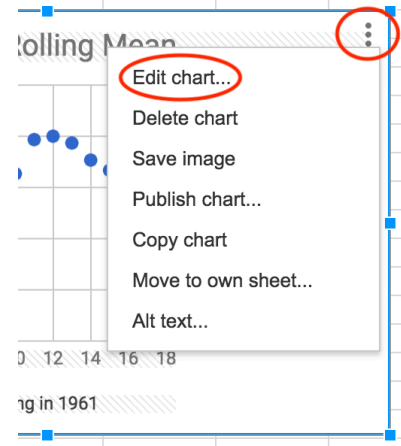
1. Select the column or columns you want to use as y-values, including the header that you want to use in your legend, which should be in the top row.
2. Note that it is okay to have blank lines in your data set (e.g. for rolling averages or missing years).
3. Click the “insert” menu and choose chart. A column or other chart will appear and the chart editor will open on the right.
4. As shown in the figure, select “Chart Type” and change it to “Scatter” and check the “Use row 1 as headers” box. This will give you a scatter chart with your y-values in it and a legend.
5. Then, as shown in the second figure, click the marker on the right edge of the “x-axis” box and select your x-values, also including their header (rows should match the rows of your y-values) and then click “ok.” This will insert the x-values onto your chart.
6. Click the “Customize” tab in the chart editor to add titles for your chart and axes. Use the dropdown under “chart title” to find the axis titles and select each axis one at a time.



Mechanics – Pro Tip for Getting Multiple Graphs in Minimal Time

The best way to get more than one graph quickly is:

- First create, fully format, and finish one graph completely, including trendlines, legends and any other features you may want.
- Then copy the entire graph (ctrl-c) and paste it (ctrl-v) to create a second copy.
- Then, on that copy, select the 3 dots in the upper right hand corner and choose “edit chart.”
- Finally, select your new y-values or make any other changes required to get your new (but already formatted) graph.



Mechanics – How to Add a Linear Trendline, Equation and R-Squared Value

Once you have a graph in Google Sheets, you can add a trendline, its equation and the r^2 -value in just a few clicks.

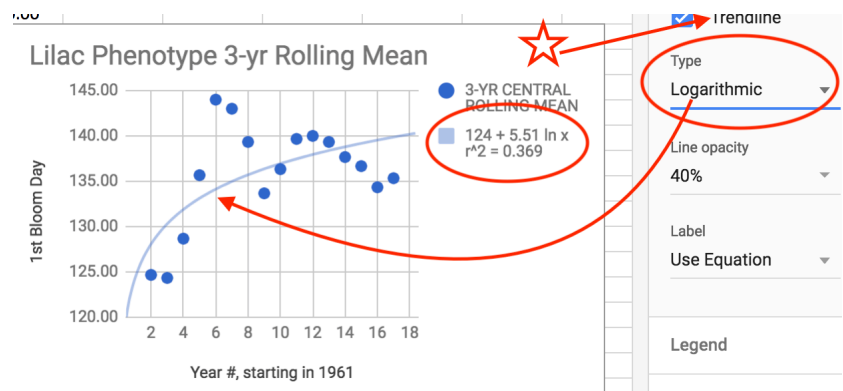
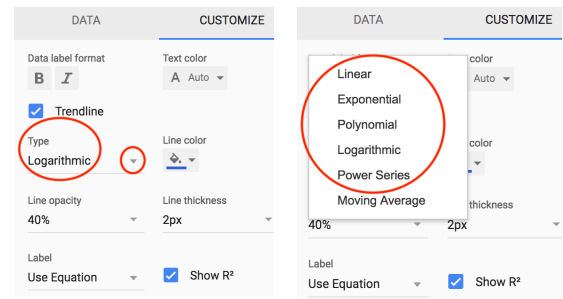
1. Click the 3-dot symbol in the upper right of the graph to get the chart editor.
2. Click on the “Customize” tab on the right and scroll down to “Series”. Click the down arrow for “Series” to expose the series options.

3. Click the boxes next to “Trendline” and “Show R²”. Then under “Label”, use the dropdown to choose “Use Equation”. This will add all three components to your graph.
4. The linear equation and r^2 -value will be shown in the legend.

Mechanics – How to Experiment with Non-Linear Trends & Tweak Format

Once you have a linear trendline, equation and r^2 -value on your chart, you can format them and experiment with other non-linear trends. Do this only if you have reason to believe the underlying phenomenon you are measuring is likely to fit another type of curve.

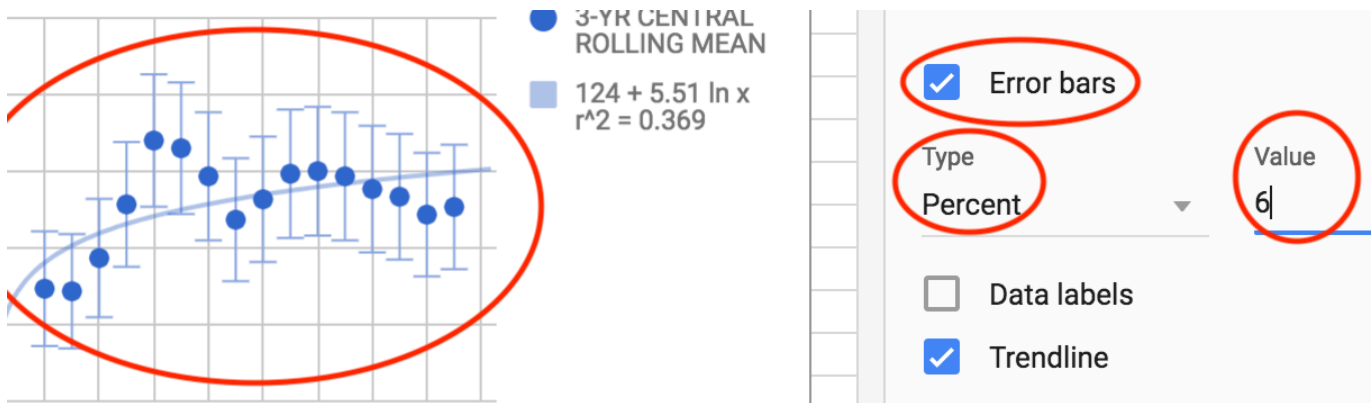
1. By clicking “Type” under the “Trendline” indicator in “Series”, you can choose among five possible trend-curve shapes. Linear is the default, but exponential, polynomial, logarithmic and power are all possible.
2. Once you choose one of these, the equation will show, along with the r^2 -value for the curve, as highlighted in the logarithmic example shown here.
3. Comparing r^2 -values can be one input to consider when deciding whether or not a particular curve shape is appropriate for a best-fit regression. That said, though, analysts should not lose site of the underlying phenomenon that is being modeled. Does it make sense that the growth rate might be increasing or slowing? The curve selected for a predictive model should make sense in light of the natural phenomena being modeled.
4. Other options for formatting trendlines on your charts are shown in the same area of the chart editor. These include line color, line opacity, line thickness, and alternative labels (instead of the equation).



Mechanics – How to Add Error Bars or Data Labels

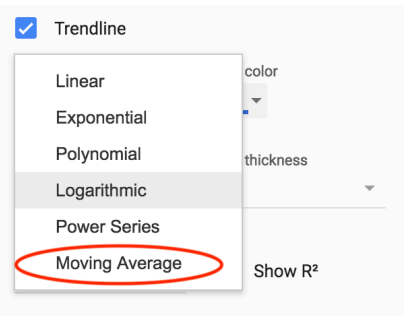
Boxes for adding rudimentary error bars and/or data labels are positioned just above the “trendline” checkbox in the chart editor. Google Sheets has only constant or percentage error bars, whereas MS Excel can add Standard Error or other options (for those who have studied the underlying statistics).

One use of error bars in Google Sheets to get a quick feel for how far from the regression curve your data points are sitting. In the example shown here, we can see that all of the data points are within 6% of the regression curve.



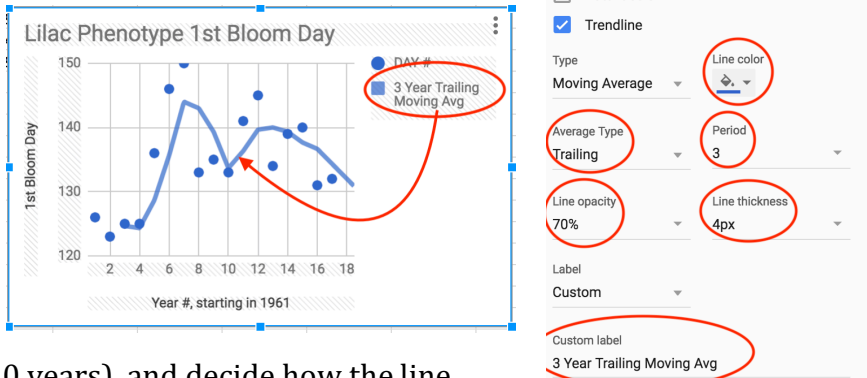
Mechanics – How to graph a Rolling Mean or Median on the Same Graph as a Scatter Plot

In Google Sheets, there are two ways to get a rolling mean onto a scatter plot along with the original data. The first gives a quick and simple view of a rolling mean, but does not provide access to the underlying data points and does not work for medians. The second provides access to the data and works for median data as well.

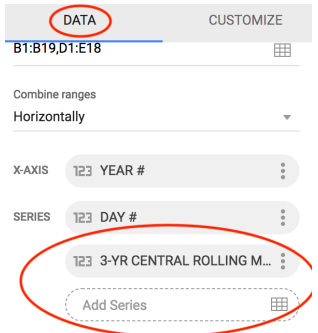


Please note that moving (rolling) averages and medians were discussed in some detail in two previous sections of this document on pages 6 and 7.

1. The quick method is to choose “moving average” from the “Types” menu of the Trendline section under “Series”. Once that has been selected, options will appear that allow you to choose a “trailing” or “central” moving average, determine how many year to average (options are available from 2 to 10 years), and decide how the line should be annotated in the legend. These options are shown in the figure.



2. A more detailed method for getting a rolling median or mean is to calculate it ourselves, as outlined in detail in the earlier sections of this document on pages 6 and 7. We can then plot it as a second series on the same chart as the original data by using the “Add Series” option under the “Data” tab, as shown on the right. Unfortunately, in Google Sheets, we cannot graph this series as a “connect the dots” scatter plot (you CAN do this in Excel), so there will be no line for visual reference. However, for a strictly visual reference, a 2-year moving average of this 2nd line provides a reasonable visual in many cases (as shown in the example below). Note that this trendline (or any other specific change) can be made for just one series by selecting that series under the “Apply To” option in the “Series” section on the “Customize” tab (also shown below).



Apply to: 3-YR CENTRAL ROLLI...

Color

Point size: 10px

Point shape: **▲ Triangle**

Error bars

Data labels

Trendline

Type: Moving Average

Line color: [Red]

Average Type: Centered

Period: 2

