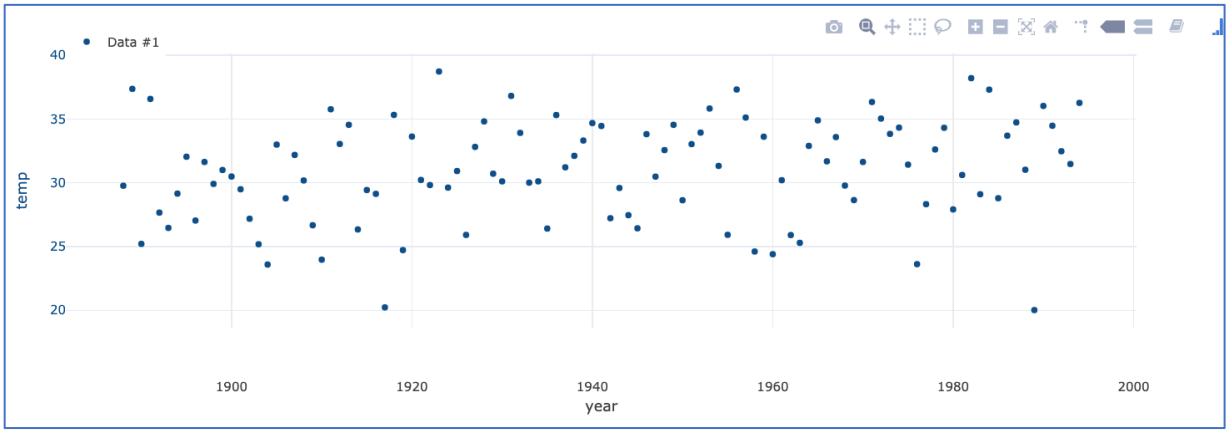
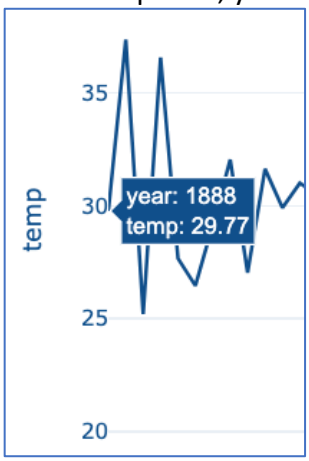
Linear Regression Example using data from Lesson 2 (STCDecTemp.txt):

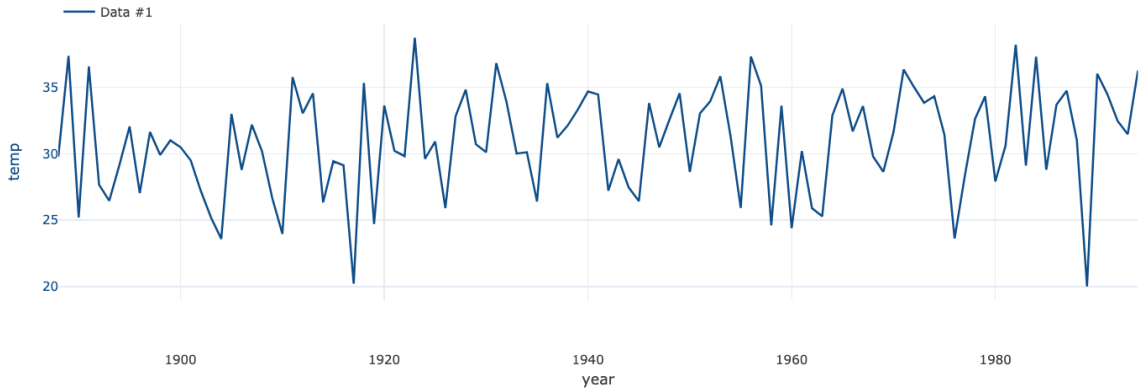


This is how the tool looks when you load the data. These data are year (x-axis) vs. mean December temperatures from State College, PA (y-axis). If you hover the mouse over the individual points, you can see the values for various years. The first and last points:



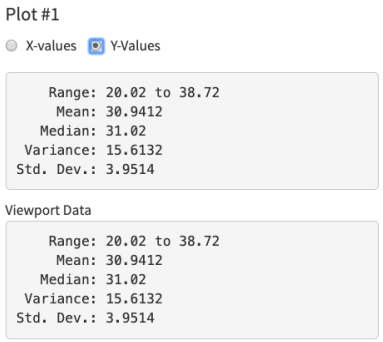
There are four tabs on the bottom of the tool. We can explore all four:

1. **Plot settings**: Control what is plotted above. Plot 1 is Blue, Plot 2 is Yellow, and Plot 3 is Green. The y-axis for plots 2 and 3 appear on the right side vs the left side. Clicking on scatter plot, line plot, or line+scatter plot just changes the type of plot. In some cases, line plots are easier to see, in others scatter plots are more useful. In our example above, it is easier to see the data with a line plot:

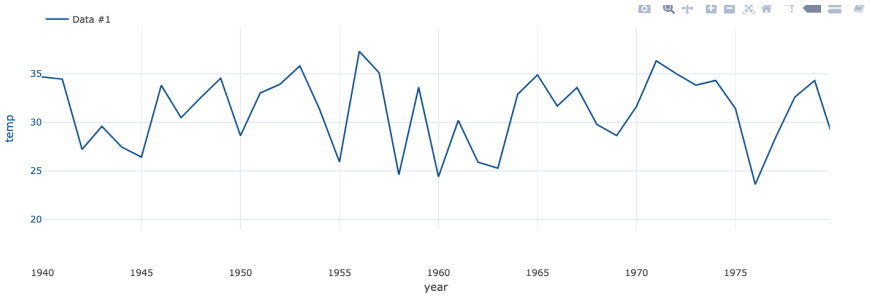


2. **Statistics**:

Shows the basic statistics of the plot above. If you zoom in on a section of the plot, you will see the statistics for just that section in the "Viewport Data" part of the tool. Make sure you click on "Y-values" to get the statistics for temperature in this example.



Here's an example where I zoomed in (click on zoom tool, and then click and drag with a mouse). I chose to zoom in on the years 1940-1980. The range, mean, median, variance, and standard deviation of the State College December temperatures during that 41 year period are shown in the now updated "Viewport Data" box below:

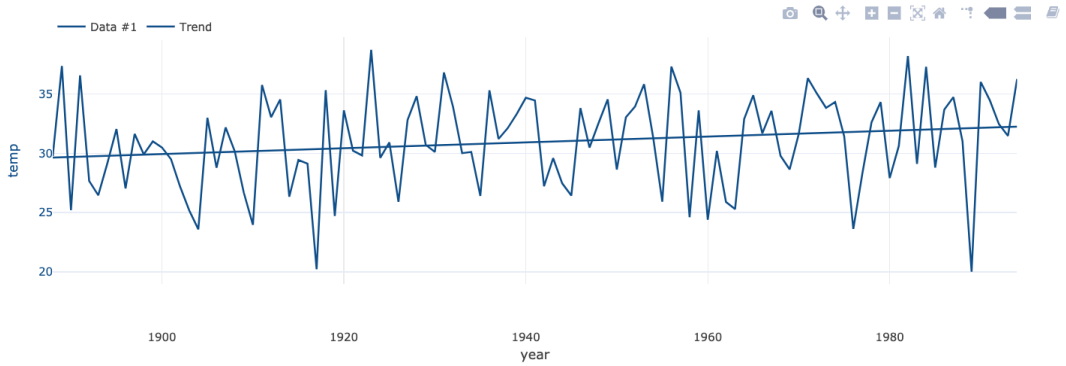


Viewport Data

```
    Range: 23.62 to 37.31
     Mean: 31.2295
   Median: 32.12
 Variance: 13.9262
Std. Dev.: 3.7318
```

Median is the value where 50% of the data are above and 50% are below this point. The mean in this case is also a fairly good measure of the center of these data. Variance and standard deviation are two different measures of how much the data varies about the mean. For the period of 1940-1980, most of the December mean temperatures in State College were within ~3.7 degrees F of 31.2F, or to put it another way, a majority of the values were between 27.5 F to 34.9 F. The actual range is the top number. The highest and lowest values between 1940 and 1980 were 37.31 and 23.62, respectively.

3. **Trend Lines**:
For trend lines, we are fitting a line to the data. We can choose all of the data, or only the zoomed in part. I have reset the axes back to all of the data and fit a line here:



Below the plot, we get some information about the line that has been fit to our data:

Plot #1

Linear Regression

All Data Viewport

Equation:

Y=0.0246*X+-16.7749

Correlation Coefficient (r):

r=0.1931

Standard Error of the Slope (Sb):

Sb=0.0122

Two-Tailed P Value (P):

P=0.0463

First, we get an equation in form of y=m*x+b, which you may remember from Algebra as the general equation that describes a line. Here, m Is the slope of the line and b is the y-offset. By definition, m is defined as the change in y divided by the change in x (rise over run), so the units of m are (units of y/units of x). In this case, since y is temperature in degrees F and x is year, m has the units of F/year. Based on the line fit to the data, we can say that there is a linear trend of 0.0246 F/year. As mentioned before, these data are from 1888-1994 (107 years), so if we multiply the slope m by the number of years N, we get m*N= 0.0246 F/year * 107 years = roughly 2.6 F change from 1888 to 1994 based on the linear trend. If I mouse over the beginning and end values on the line and subtract the 2 values, I also get roughly 2.6F!

The Y-intercept, b, has the units of the y-axis, degrees F. It is not really physically meaningful in this example, as we do not have any x values that are 0, but it would be useful if we wanted to use our linear trend, which is just a very simple model(!), to predict other values of y given x. For instance, if we wanted to predict the value of State College December mean temperatures for 2020, we could put the year 2020 in our equation above and get:

Y=mx+b=0.0246 *2020 + (-16.7749)=32.9F

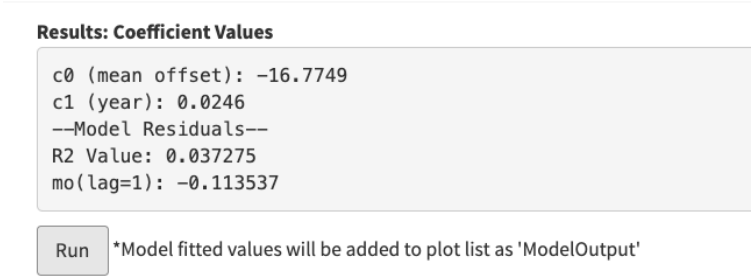
The rest of the values in the box tell us a little bit about the quality of linear fit, or linear model, as it pertains to our data.

The correlation coefficient, r, tells us how closely related the x and y values are. Perfectly related values have an r-value of 1 or -1, depending on whether the increase in x yields an increase in y or a decrease in y. Values close to 0 indicate that there is little relationship between the two set of values. Here, we see that increase in year does yield a slight increase in State College December temperatures. We will test later to see if that relationship is statistically significant.

Standard Error of the slope, Sb, gives us the 95% confidence interval our linear model. It has the same units as slope. In general, we consider the 95% confidence interval to be m +/- 2*Sb. In this case, it would be 0.0246 +/- 2*0.0122= 0.0002 F - 0.049 F. Since this range (just barely) does not include 0 F, it is one measure that the increase in State College December mean temperatures is significant.

We'll cover p-value a bit later.

4. **Regression Model**:
Here's where we actual make the linear fit model. If we chose 'year' in Model Parameters and 'temp' in Target Observation, we get this in the results box:



We noticed that the coefficients in our simple linear model exactly match the line we fit in the previous tab, where c0=b and c1=m. We have again made a simple linear model which here has the form y=c0+c1*x .

The R2 (or R^2) value is 0.037. This value is telling us what fraction of the overall variability our model has captured. In this case, it's very little, just 3.7% of the overall variability. You can see that in the fact that there is a great deal of year-to-year variation that is not described by our linear trend.

Finally, the "mo(lag=1)", also called rho or ρ elsewhere in the text, is the autocorrelation value, which we will discuss later.