Exploring Weather Trends

Paul Sterk
Data Analyst Nanodegree
3/21/2020

Summary

In this project, we will analyze local and global temperature data and compare the temperature trends where you live to overall global temperature trends.

Steps

The goal of this project is to create a visualization and prepare a write up describing the similarities and differences between global temperature trends and temperature trends in the closest big city to where you live. In my case, that is Sacramento.

Below is an outline of steps taken to prepare the data to be visualized in the char:

Extracted the Data

- Used SQL to create a result set of global daily average temperatures (°C) for the city Sacramento.
- Exported the results as a CSV file to my local hard drive and renamed the file as “sac_annual_avg_temp.csv”.
- Used SQL to created a result set of global daily average temperatures (°C)
- Exported the results as a CSV file to my local hard drive as “global_avg_temp.csv”

Opened up the CSV

- Logged into Google Sheets using my Google account.
- Imported city CSV file into Google Sheets and saved a new spreadsheet

- Imported world (global) CSV file into Google sheets and inserted into the previously created spreadsheet. The results for Sacramento on the left and global on the right.
- Calculated the 7-day moving average for temperatures with the AVERAGE(dataset) function for both Sacramento (1849-2013) and global datasets (1750-2015).
- Calculated overall average and standard deviation for both datasets.

Created a Line Chart

Compared my city's temperatures with the global temperatures. Plotted the moving average rather than the yearly averages in order to smooth out the lines, making trends more observable a spreadsheet).

Key considerations when deciding how to visualize the trends:

Considerations for all Use Cases

- Principal of Least Effort.: the visualization should also intuitively convey the essential point of the chart with the least amount of effort on the part of the observer.
- The important point is to convey the essential essence of the data with introducing visual "noise".
- One approach is to try several different chart types and see which one best "fits" the data.
    - If the data looks like a line (low variance on the y-axis), use a line graph.
    - If the data looks has points spread out (high variance) in both x and y dimensions, use a scatter plot chart.
    - If the data depicts not points, but ranges of values on the y-axis, use an area chart
    - If the data depicts parts of a whole, use a pie chart.
    - If the data has geographic coordinates, use a geo-based chart.
- In general, use the chart which best conveys the underlying structure or form of the data.

Considerations for this Use Case

- There are 267 data points in the global data set, 164 data points for Sacramento.
- Since the data points have fairly low variance with respect to the y-axis, a scatter plot will show a clustered series of points similar to a line. The result will be a “fat” line of connected points. If the shape is a line, why not just depict it as such.
- Since we are interested in answering questions such as: “are average annual temperatures rising or falling, using a line is an easy way to depict whether the values is going up over time (positive slope) or going down (negative slope).
- We could have displayed a series of points, but it would take more effort on the user to determine if the slope is positive or negative.
- Since this is a time-series dataset, it is important for the graph to emphasize the sequential, connected nature of the data. One year follows another in order. Seems obvious, but worth mentioning, IMO.
- Therefore, I created a line plot for both datasets using this functions feature in Sheets. Results shown below.

Observations

Similarities and differences between the world averages and your city's averages, as well as overall trends.

Sacramento Compared to Global: 1849 - 2015 - Summary

Sacramento's average annual temperature is always higher than the global average. Summary statics below:

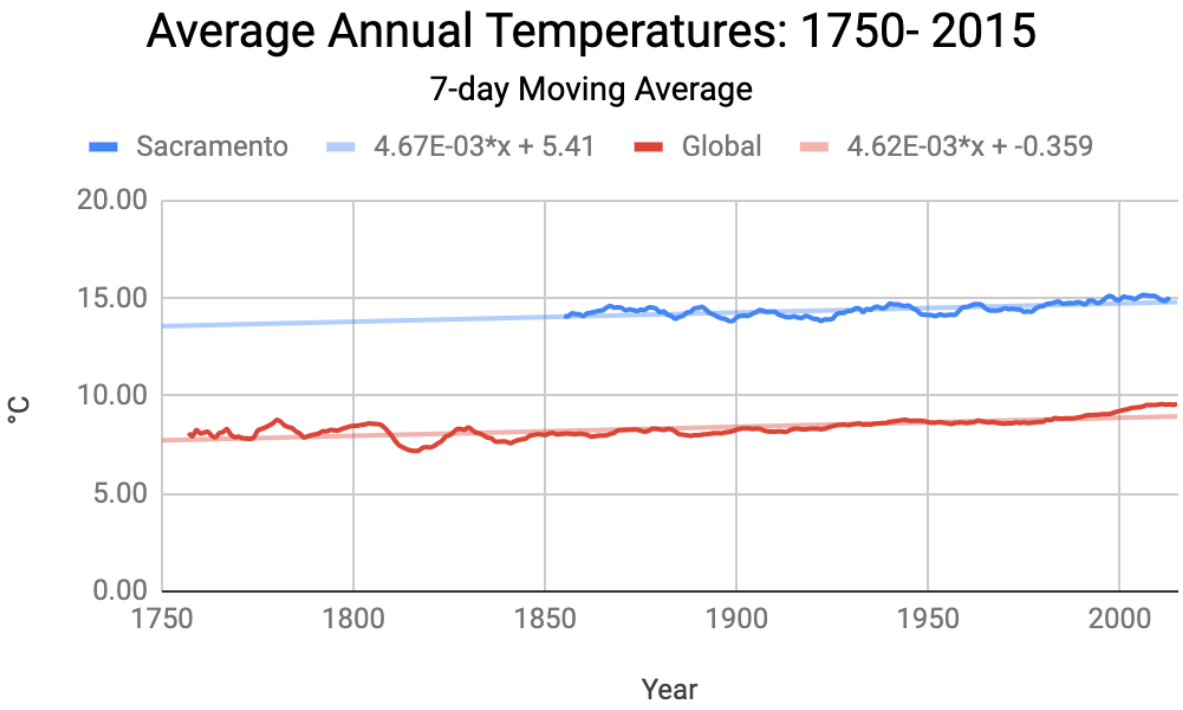
Location	Average Annual Temperature °C	Standard Deviation
Sacramento	14.45	0.50
Global	8.37	0.58

Table 1

The difference between the two averages is 14.45 - 8.37 = 6.08. Given that this difference is so much higher than the standard deviation of the Global temperatures (0.58), we can conclude that Sacramento is significantly warmer than the global average. Given that Sacramento has a hot, Mediterranean climate, this result is not that surprising. It depends on what other cities are used in the global averages and in what climate zones these cities are located.

Also, since these data points are annual averages, we would probably expected the standard deviation to be low as significant, short-term climate changes are relatively rare. The low SD values are in line with this intuition.

Sacramento Compared to Global: 1750 - 2015 - Changes



Above is a line plot of the 7-day moving average annual temperatures for both datasets. The chart also contains trend lines with respective equations. The difference between the two lines appears to be consistent with a slight narrowing between the two from 2000-2015. Both datasets have a similar, positive slope indicating a gradual rise in 7-day moving average annual temperatures.

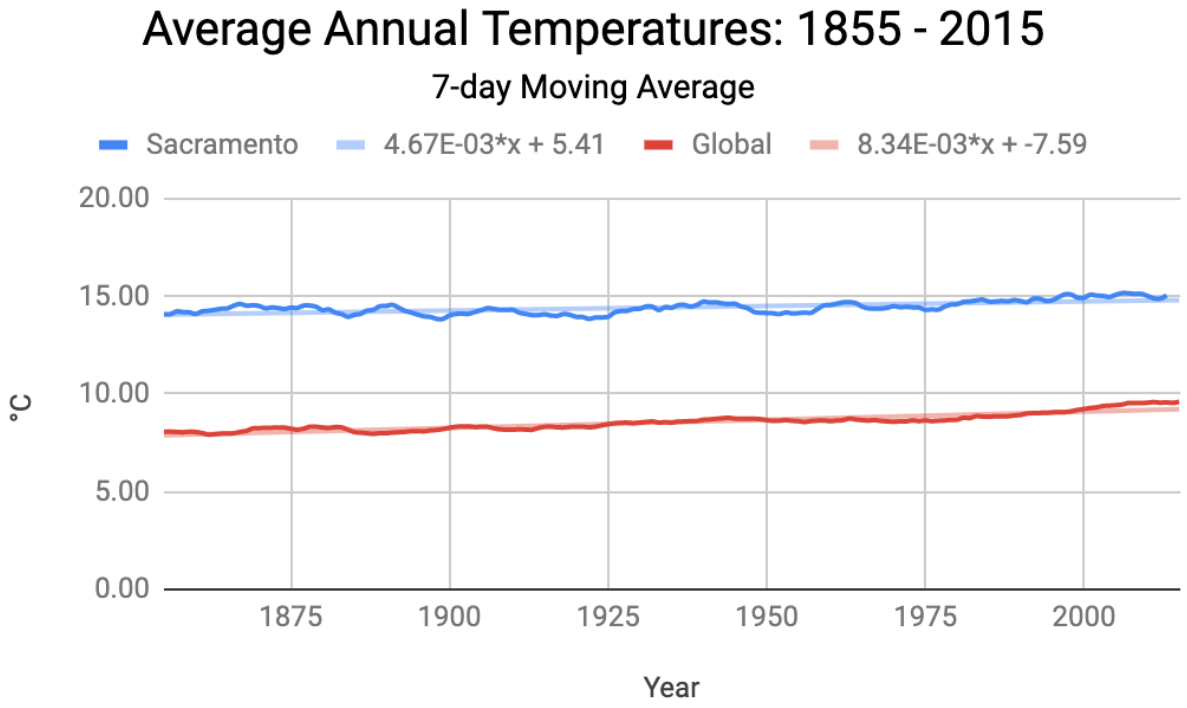
Sacramento: 4.67E-03
Global: 4.62E-03

An informal inspection of these two numbers suggests the rate of change is very similar.

While a positive slope does indicate an increase, it is 0.00462 of degree for the global dataset. *Is this this significant enough to conclude that average temperatures are rising?* Could it just be "noise" in the data and average temperatures are flat. The intuition is that this will require further statistical tests which are out-of-scope for this assignment.

One thing to note is that because the state California was founded in 1849, we have temperature data starting from that year. Let's look at the trend lines starting in this year to see if any additional conclusions can be reached.

Sacramento Compared to Global: 1855 - 2015 - Changes



Over the years 1849 - 2015, both trend lines have a slight upward (positive) slope. Slight defined as a slope close to zero. Trend lines with equations are also included:

	Slope (rate of increase)
Sacramento:	4.67E-03
Global:	8.34E-03

By visual inspections, Sacramento has higher variation in the data than the global averages. Meaning the distance between the light colored trend line and the line data is greater. One explanation is that global data has measurements from cities around world. This greater variety of locations could result in "smoothing out" of variation in annual averages.

Overall Trend

Since the slope of both trend lines is positive, this might indicate that both Sacramento and the world are getting hotter. Again, 0.00834 of degree change is not in itself conclusive that the world is getting hotter without further statistical tests. All you can really say is a casual inspection might suggest it.

What is very interesting is that the rate of increase (slope) for Global 7-day moving averages has almost *doubled* from 4.62E-03 to 7.77E-03:

Years	Slope (rate of increase)
1750 - 2015:	4.62E-03
1855 - 2015:	7.77E-03

Is this significant difference? It might be, but again more statistical tests should be applied.

As an aside, it is interesting to note that 1855 could be considered an early period of the industrial revolution with increasing amounts of human created, green-house ages (e.g., coal-fired plants). However, there is not enough data in this set to make any definitive conclusion.

It does point out even with the same dataset, difference "slices" of the data can lead to different conclusions.

Is the Trend Consistent?

The overall trend line has a positive slope indicating an overall increase. This could be evidence of the data being consistent. However, there have been shorter periods of declining averages. Example, from about 1880 to 1890 and from about 1962 to 1976.

However, the question, “Is the Trend Consistent” begs another: what is the definition of “consistent”. We can take a cursory, visual inspection of the at and draw some conclusions. Nevertheless, we should exercise caution when making these causal conclusions. How do we really decide that the line data is consistent or inconsistent?

This seems to prompt the need to use a standard statistical measurement to answer this question. Since this is out-of-scope for this assignment, we can explore this question in a future study.