Weather Trends Exploration

Udacity Data Analytics

09/12/2018

By Roberto Soriano

The project

The purpose of the project is to analyze the global temperature data and compare the temperature trends of the closest major city to overall global temperature trends. The database provided contained a table for major cities in the world, their historic temperature averages (in Celsius) per year (city_data). Another table the listed all the names of the major cities used (city_list). Also, a last table for the global temperature averages from 1750-2015 (global_data). The data was to be visualized and then digested, to find out what the story behind the numbers was.

Methods

First I wanted to take a peak at what was hidden inside the database. I ran these queries to observe the data I was going to be working with:

```
SELECT * from city_data
SELECT * from global_data
```

To find which big or “major” city was closest to me, though I already had an idea of what the answer would be:

```
SELECT * from city_list
WHERE country = 'United States'
```

After I saw that Boston was listed (obviously), I ran the query below to display the years, and their respective temperatures, for the city of Boston only:

```
SELECT * from city_data
WHERE city = 'Boston'
```

After having a better idea what the data was like, I realized the CSV file I was to export from the database should contain the following columns:

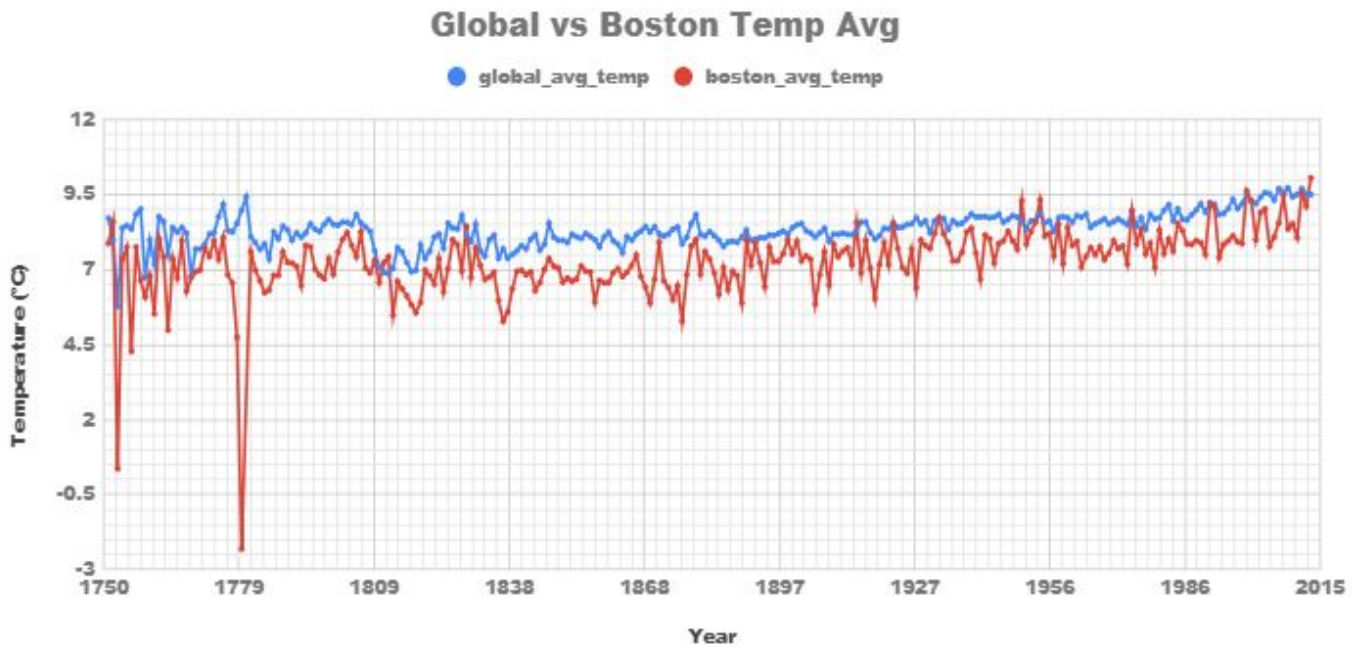
Year (Where it matches Boston only)	global.avg_temp	boston.avg_temp

I knew I had to SQL to manipulate the data before exporting it, thus making it easier for working with it afterwards. This part took some time. At first, I exported two separate sets of data, for global_data.avg_temp and city_data.avg_temp. However, I knew this wasn't going to cut it, because I would then need to join/merge tables "manually" using the spreadsheets software, which wouldn't be time efficient at all. After some research and refreshers on the SQL language, the most neat and succinct query I could use to get all the data I needed in one set, with only the years that corresponded to temperatures on both table:

```
ALTER TABLE global_data
RENAME COLUMN avg_temp to global_avg_temp
ALTER TABLE city_data
RENAME COLUMN avg_temp to boston_avg_temp

SELECT global_data.year, global_data.global_avg_temp, boston_avg_temp
FROM global_data INNER JOIN city_data
ON global_data.year=city_data.year
WHERE city = 'Boston';
-Exported and saved as globalvsboston temps (CSV file)-
```

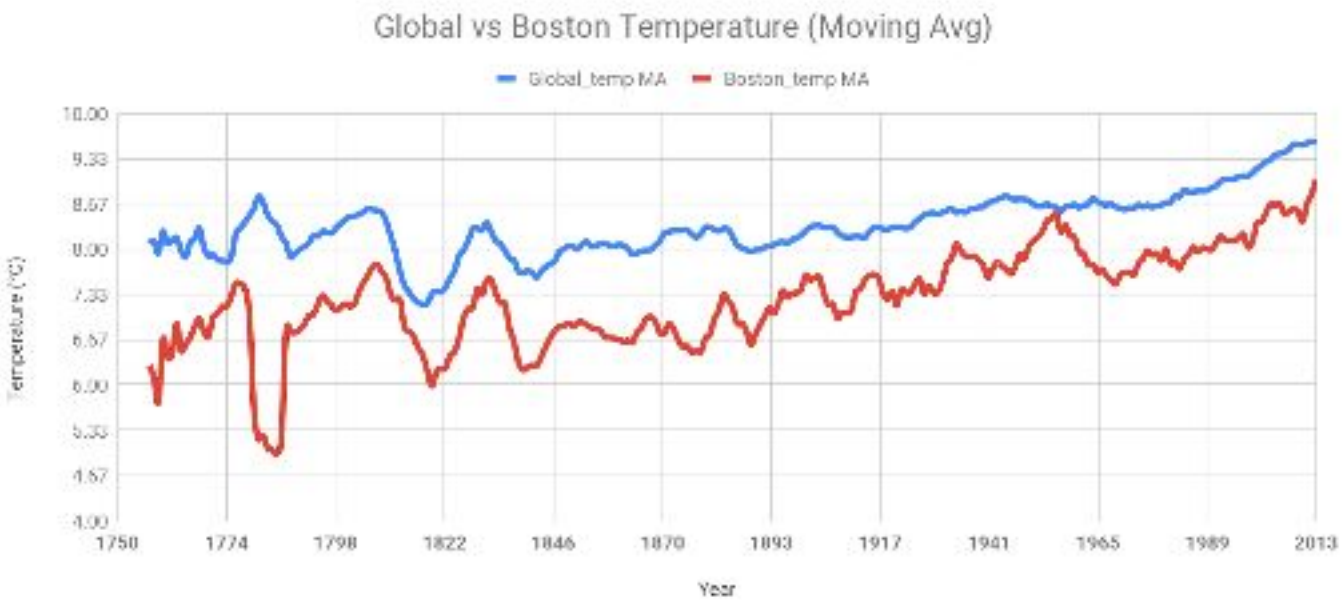
With the dataset generated above, I proceeded to create a line graph. I believe this would be the most simple and straightforward way to display the findings. The x axis is for the "year", starting 1750, since this is the first year in common for the temperatures, in Celsius, on y axis.



The blue line graphs the global temperature averages. The red line graphs the Boston temperature averages. However, these lines looks unnecessarily sloppy and volatile. It is harder to tell the story, or analyze the data, when the lines are not viewer friendly. To smooth them out, we then use the moving averages out of the averages. I created two new columns, Global_temp MA and Boston_temp MA, where MA = moving average).I used a 7 year window for my calculations of the moving averages. The formula I used in the spreadsheets is: =AVERAGE(firstvalue locations:lastvalue location). For the first instance, the exact formula looked like =AVERAGE(B2:B8).
The final table looked something like this :

Year (Where it matches Boston only)	global.avg_temp	boston.avg_temp	Global_temp MA	Boston_temp MA
1750	8.72	7.88	-	-
1751	7.98	8.6	-	-
1752	5.78	0.36	-	-
1753	8.39	7.35	-	-
1754	8.47	7.75	-	-
1755	8.36	4.28	-	-
1756	8.85	7.76	8.078571429	6.282857143
1757	9.02	6.65	8.121428571	6.107142857

The final graph displays the global and Boston's historic temperature moving averages:



Observations

By just glancing at this graph, we can quickly discern many insights of Boston's temperature averages that "jump-out of the page":

(1) Boston's historic temperature averages are significantly lower than the global averages. Among other things, means that Boston is colder than your average Joe!
(2) Boston's temperature averages are more volatile. There are more ups and down on the red line. This means Boston temperature changes are more drastic from year to year.
(3) Despite of the ups and down, the global is warming up! Both lines have a positive slope, average temperatures show to have increased over time.
(4) Comparing the global temperature averages as we have done on this graph could be misleading! All of the cities we could choose from were "major" cities. One key

characteristic of major cities is that they are near the water (for the most part). Why could this be a problem? In Boston's case, for example, temperature averages might be lower than that of its surrounding cities because of it being a coastal city. Of course, this could go all the way around for cities which are further away from the poles or in tropical areas.