Udacity : Data Analyst Nanodegree

Project: Exploring Weather Trends

Comparison in Weather Trends Around the Globe

Introduction

The project assignment is to compare weather data between the global average temperature data and data from a city closest to where I reside. Assigned tasks are the following:

- Extract relevant data from a database using SQL.
- Open up the CSV, and prep the data for visualization.
- Create a visualization of the data.

- Report on findings from the data.

Step 1: Extract the Data

The first step is to extract the data from the given tables on Udacity website. Given that I reside in Pittsburgh, Pennsylvania in the United States, I started looking for Pittsburgh or a city nearby. I initially started by browsing through the `city_list` table with the following query:

```
SELECT * FROM city_list;
```

The table displayed the following data:

Output 345 results Download CSV

city	country
Abidjan	Côte D'Ivoire
Abu Dhabi	United Arab Emirates
Abuja	Nigeria
Accra	Ghana
Adana	Turkey
Adelaide	Australia
Agra	India

I saw in the table that each data point in `city` and `country` columns are spelled in capital letters. Thus I tried to directly query for Pittsburgh:

```
SELECT * FROM city_list WHERE city = 'Pittsburgh';
```

Unfortunately I had no results, so I decided to see what cities are available in United States:

```
SELECT * FROM city_list WHERE country = 'United States';
```

Success! I get a list of cities with available data. Browsing through the table I see a city nearby that I could use; Columbus, Ohio. I go ahead and extract Columbus's average temperature data from in the `city_data` table:

```
SELECT * FROM city_data WHERE city = 'Columbus';
```

The query successfully returned average temperature data for Columbus and thus I saved it as a 'Columbus.csv' file.

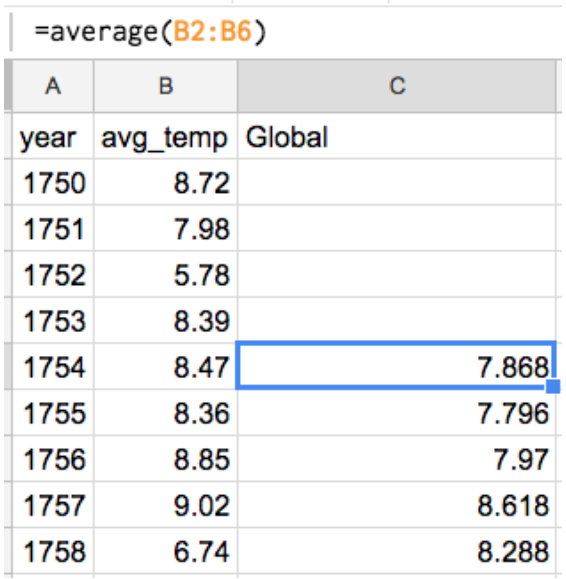
The same process goes for extracting the global data by querying the `global_data` table and saving it as a 'Global.csv' file.

```
SELECT * FROM global_data
```

Step 2: Preparing the Data

Now that both 'Columbus.csv' and 'Global.csv' are exported, the next task is to prepare the data for visualization. I chose to use Google Sheets for this task, as I am most familiar it from using it at work. Steps taken to prepare the data are shown below:

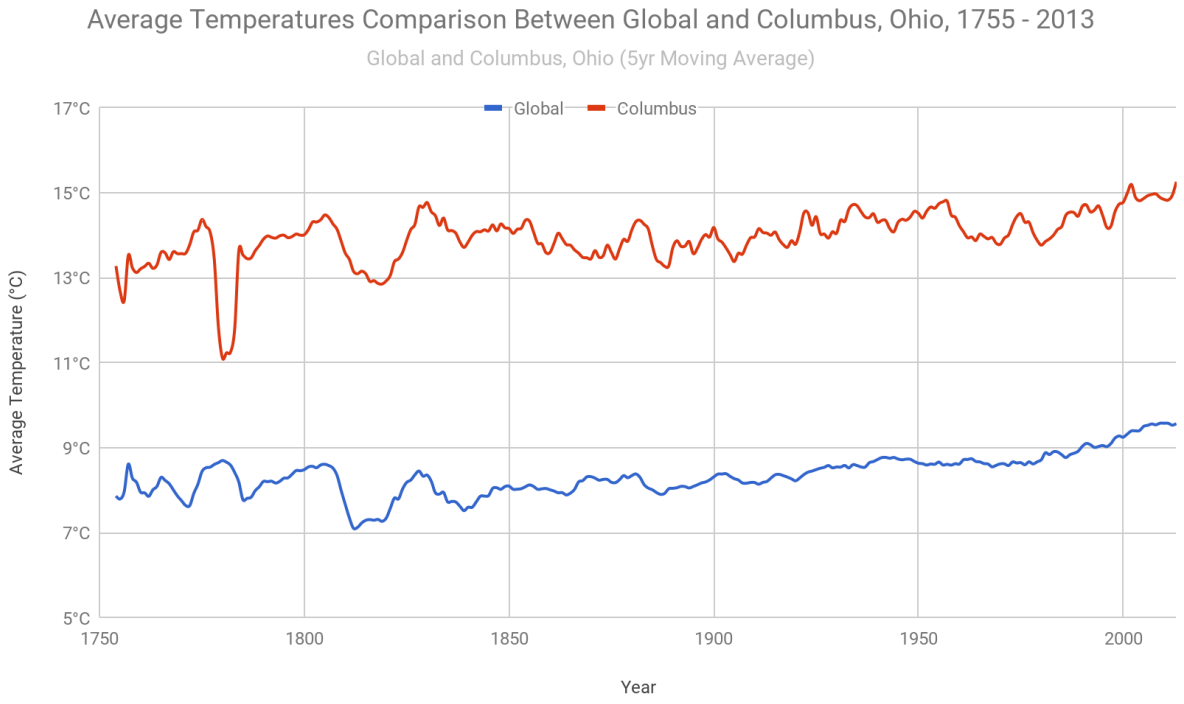
1) Import both csv files into Google Sheets.
2) To display a smoother reading of the data, I chose to use a 5 year moving average. This was done using the `AVERAGE()` formula by selecting 5 consecutive years.



3) The relevant starting year for each data set was different, so I started the moving average from 1750, as it is good cut off point for setting a date range.

Step 3: Visualizing the Data

Using the data organized in step 2, I used Google Sheets' chart function to build the line chart.



Note that the moving average required five years from 1750, so the actual data shown begins from 1755.

Step 4: Analyze the Data

Some findings from the data:

1) Columbus' yearly temperature is consistently warmer than the global data. The average yearly temperature in Columbus, Ohio is on average **5.62** degrees celsius warmer than the global temperature[1]. I assume this is due to the global average

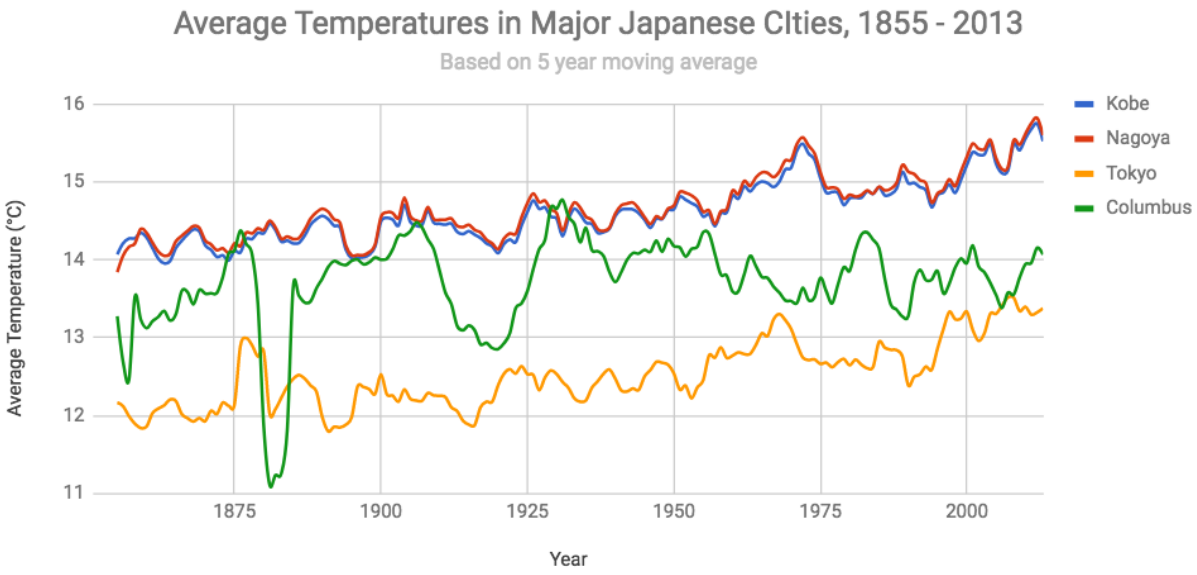
[1] Calculated using AVERAGE() function on the avg_temp columns.

incorporating data from the Arctic and Antarctic regions, as well as data from high elevation areas.

2) The movement of average temperature showed a moderate correlation between the global data and Columbus data. I calculated the correlation coefficient using the Pearson correlation function in Google Sheets[2]. The data showed that the correlation was **0.55**, showing a moderate uphill relationship. This is supported by the chart, as it shows a gradual upward trend both globally and in Columbus.
3) I noticed a large dip in the average data for Columbus, Ohio around the period of 1780-1782. At first I wondered if this was faulty data, but I did search around online and found some interesting historical stories about historically cold winters around that time.
4) Global data shows a gradual uphill trend in increasing average annual temperature. Using the 5 year moving average, I calculated the percent difference each year by calculating `C1=(B1-A1)/A1` where A1 is the first moving average, B1 is the following year's moving average, and C1 is the percentage difference. Applying this to all the moving average temperature, the calculation showed that the global yearly temperature showed an average increase of **0.08%**.

[2] Calculated using CORREL() function on the avg_temp columns.

Bonus: Comparing Japanese Cities to Columbus



Since I used to live in Japan, I decided to compare Columbus, Ohio to some of the major cities in Japan. The moving average of Columbus generally sits between Kobe & Nagoya and Tokyo. Kobe and Nagoya has a very high correlation (**0.99**). Compared to that, correlation between Kobe and Tokyo is **0.68**. It is interesting to note that the temperature dip in the Columbus data between 1779-1783 (as reported in comment #3 above) is especially pronounced when comparing to data from the other cities.