Weather Trends

Steps Taken

Accessing Data With SQL

In this exercise, three tables are provided in a SQL database:

- `city_list`: made of columns `city`, and `country`. For this exercise I'm searching the city I live in; Madrid. The query used to do this is:

  ```
  select * from city_list where city = 'Madrid';
  ```

  With this, I'm making sure that my city will (or should) be listed in the following table: `city_data`.

- `city_data`: composed of columns `year`, `city`, country and avg_temp. In my scenario, I want to filter the results out by `city`, and I'm only interested on `year` and `avg_temp`:

  ```
  select year, avg_temp from city_data where city = 'Madrid' and
  country = 'Spain';
  ```

  While the country may not be needed at all, it's safer to just add it to the conditions, as some other hispanic countries share city names with Spain. More important than this are the field selections, which will filter the signal (valuable data) out of the noise (invaluable columns). The results of this query will be used for the first .csv export.

- `global_data`: in this case, the table is pretty simple and already contains all the data we want; `year` and `avg_temp`, so the query is simple:

  ```
  select * from global_data;
  ```

  The results of this query will be used for the second .csv export.

Manipulating Extracted Data

Once I got the 2 main `.csv` files (one for the local data, and the other one for the global values), I imported both into a Google Spreadsheet, each one into a dedicated sheet. Properly formatting the imported columns (`year` and `avg_temp`) as numbers.
Then, I added a 3rd column to each of the sheets, which contains the 10 year-window moving average of the `avg_temp` column. This was done as shown in the picture below:

C11 | fx | =AVERAGE(B2:B11)

	A	B	C
1	year	avg_temp	10y_global_avg
2	1750	8.72	
3	1751	7.98	
4	1752	5.78	
5	1753	8.39	
6	1754	8.47	
7	1755	8.36	
8	1756	8.85	
9	1757	9.02	
10	1758	6.74	8.03 ×
11	1759	7.99	=AVERAGE(B2:B11)

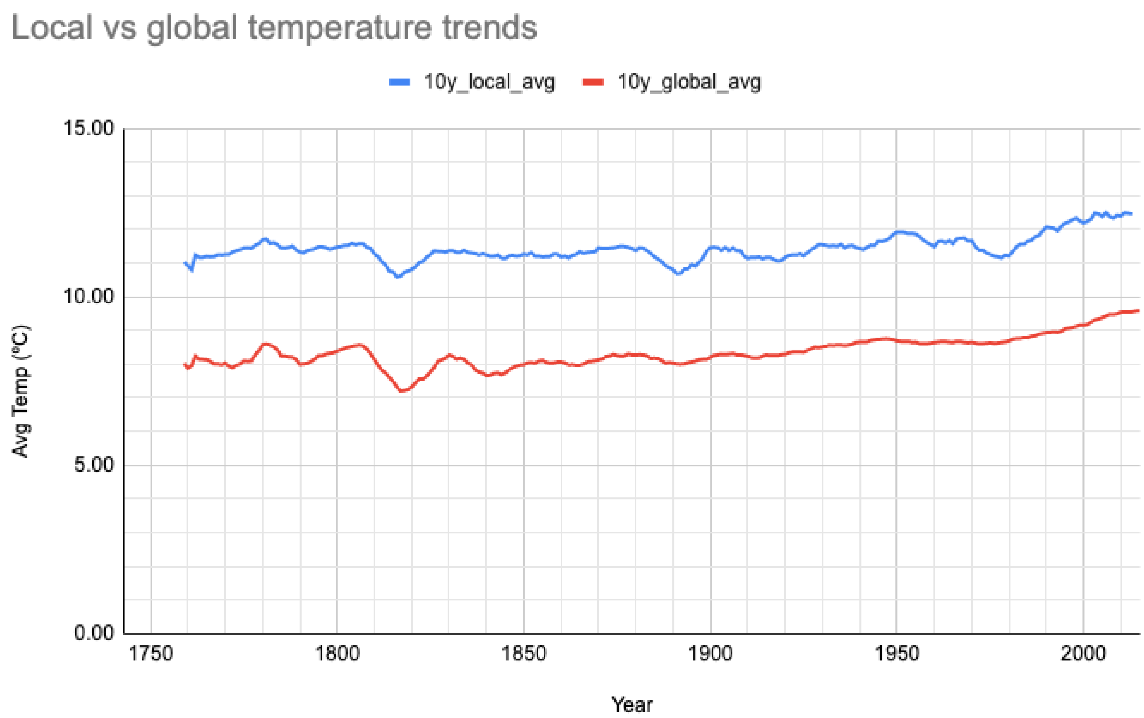
C12 | fx | =AVERAGE(B3:B12)

	A	B	C
1	year	avg_temp	10y_global_avg
2	1750	8.72	
3	1751	7.98	
4	1752	5.78	
5	1753	8.39	
6	1754	8.47	
7	1755	8.36	
8	1756	8.85	
9	1757	9.02	
10	1758	6.74	
11	1759	7.99	7.88 × 8.03
12	1760	7.19	=AVERAGE(B3:B12)

This formula was applied for every temperature value of each sheet, correspondingly (except for the initial 10 temperature values). The decision on the 10 year window for the moving average wasn't arbitrary. Since we have data from 1750 to more or less 2015, I thought it was a good balance between resolution and precision for the later graphic interpretation.

Visualizing Manipulated Data

At this point, I already have the data I need to plot a graphic: `year` (for the X axis), `10y_global_avg` (1st series) and `10y_local_avg` (2nd series). In the resulting chart, we can observe as the legend indicates, the evolution of the local average temperature in blue, and the global average in red:



Conclusions

Observing the chart there are some takeaways already:

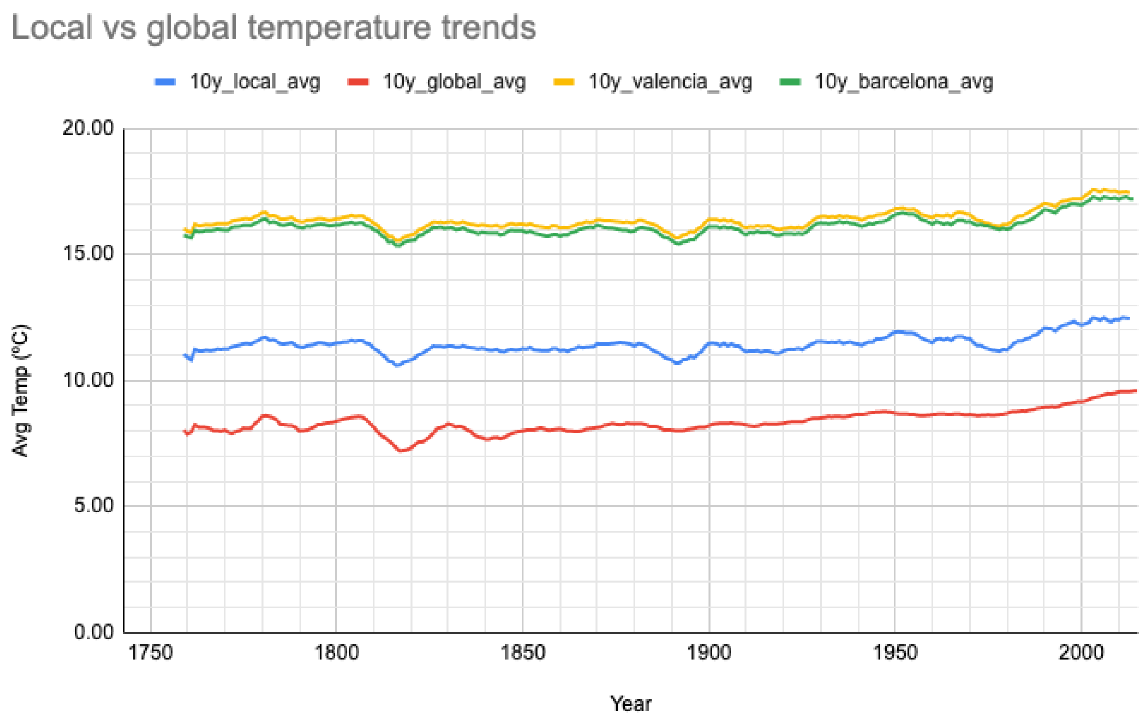
- The temperature in Madrid has always been greater than the global average in around 3ºC
- Both local and global temperatures have been rising since the beginning of the dataset
- The global temperature has increased around 1.5ºC
- The local temperature has increased around 1.5ºC as well
- Both averages seem to keep this trend in the future
- Looking at the local temperature in depth, we can observe how there is an ~80 year period when temperatures decrease almost 1ºC (blue line values around 1815, 1890 and 1975).
- From the point above, one could expect that the temperatures in 2060 in Madrid could be one degree cooler than some years earlier

Further Study

The latest 2 bullet points from the conclusions awakened a curiosity in me. So I wanted to check how the temperatures were in other cities from my country during these years too. Thus, to know which cities from my country were listed in the dataset, I ran the following query in the city_list dataset:

```
select * from city_list where country = 'Spain';
```

I got Valencia and Barcelona. After that, it was just about repeating the same process for the values of the new cities, and add them into the chart (orange for Valencia, green for Barcelona):



Effectively, we can see the same temperature decrease in these specific years. Additionally, we can see that their average temperature is way greater than both global and Madrid's. The reason behind this is that they are located by the sea, which keeps temperatures more regular over the year. A first, naive conclusion would be just saying that overall these cities are warmer. However, during the months of Summer, one can expect cooler temperatures there than in Madrid.

Another interesting metric to calculate is whether there's a clear correlation between the variables under study or not. This is done by calculating Pearson's correlation coefficient. Here were the numbers for each case:

- Global: 0.6227745063
- Madrid: 0.3826015337
- Valencia: 0.4152811533
- Barcelona: 0.4154101393

In all of the scenarios we could tell there's a slight correlation between time and temperature. The greater the year, the greater the temperature. Now, when it comes to Barcelona and Valencia, both show very similar values. This is not a coincidence, since these cities are geographically under the same conditions: both are cities next to the Mediterranean sea, and we could say they are somewhat close next to each other.

NOTE: in all of the graphs shown above, data before 1759 has been removed due to missing values in the dataset.