# Exploring Weather Trends

## Summary

In this project, you will analyse local and global temperature data and compare the temperature trends from the city of **Berlin, Germany** to overall global temperature trends.

## Goal

Create a visualization and prepare a write up describing the similarities and differences between global temperature trends and temperature trends in **Berlin, Germany.**

## Steps

1. **Extract the data**

I extracted the data from the workspace using the following query:

```
SELECT city.year
            , city.city
    , city.country
    , city.avg_temp as city_avg_temp
    , global.avg_temp as global_avg_temp
FROM city_data as city
INNER JOIN global_data as global
on city.year = global.year
WHERE city.country = 'Germany'
AND city.city = 'Berlin'
AND city.avg_temp IS NOT NULL;
```

Since the city data and the global data will be evaluated together, I decided to use one single SQL query to retrieve both, already joined by the year.
Having in mind that the expected analysis is a comparison among two temperatures I excluded all records that had no data. In the case of the city average missing is excluded by using the where clause to filter null values. To exclude the missing average values in the global data I used an inner join, excluding the records belonging to those years entirely.
The reason why I decided to exclude them instead of using a technique to fill in missing data as replacing it with the mean, is because the incomplete years were the first 7 years of the dataset and not years in between, so it would affect less the value to exclude them than to replace them.

After executing the query in the workspace, I downloaded the result set as a csv file.

2. **Open up the CSV and calculate the moving average**

I decided to do the analysis using **Excel**.
I opened the results of the query and calculated the moving average using different intervals to test which one would fit best the purpose: 5, 10, 15 and 20 years.
I calculated the moving average by using the AVERAGE formula and indicating the different intervals according to the amount of years I needed.

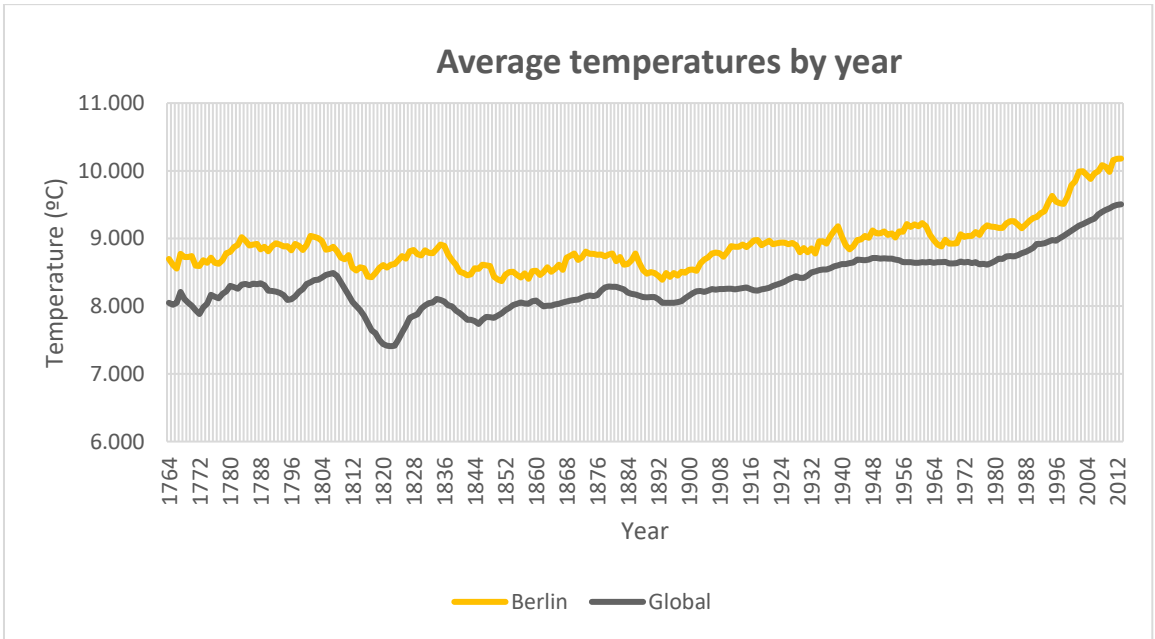
In the following snapshot the formula used for the first cell of every moving average can be seen:

| year | city | country | city_avg_temp | 5-years MA | 10-years MA | 15-years MA | 20-years MA | global_avg_temp | 5-years MA |
|---|---|---|---|---|---|---|---|---|---|
| 1750 | Berlin | Germany | 9.83 | | | | | 8.72 | |
| 1751 | Berlin | Germany | 9.75 | | | | | 7.98 | |
| 1752 | Berlin | Germany | 4.84 | | | | | 5.78 | |
| 1753 | Berlin | Germany | 8.72 | | | | | 8.39 | |
| 1754 | Berlin | Germany | 8.49 | 8.326 | =AVERAGE(D2:D6) | | | 8.47 | 7.868 |
| 1755 | Berlin | Germany | 8.26 | 8.012 | | | | 8.36 | 7.796 |
| 1756 | Berlin | Germany | 9.62 | 7.986 | | | | 8.85 | 7.970 |
| 1757 | Berlin | Germany | 9.15 | 8.848 | | | | 9.02 | 8.618 |
| 1758 | Berlin | Germany | 8.25 | 8.754 | | | | 6.74 | 8.288 |
| 1759 | Berlin | Germany | 9.04 | 8.864 | 8.595 | =AVERAGE(D2:D11) | | 7.99 | 8.192 |
| 1760 | Berlin | Germany | 8.99 | 9.010 | 8.511 | | | 7.19 | 7.958 |
| 1761 | Berlin | Germany | 9.47 | 8.980 | 8.483 | | | 8.77 | 7.942 |
| 1762 | Berlin | Germany | 8.53 | 8.856 | 8.852 | | | 8.61 | 7.860 |
| 1763 | Berlin | Germany | 8.62 | 8.930 | 8.842 | | | 7.5 | 8.012 |
| 1764 | Berlin | Germany | 8.91 | 8.904 | 8.884 | 8.698 | =AVERAGE(D2:D16) | 8.4 | 8.094 |
| 1765 | Berlin | Germany | 8.54 | 8.814 | 8.912 | 8.612 | | 8.25 | 8.306 |
| 1766 | Berlin | Germany | 8.87 | 8.694 | 8.837 | 8.553 | | 8.41 | 8.234 |
| 1767 | Berlin | Germany | 8.14 | 8.616 | 8.736 | 8.773 | | 8.22 | 8.156 |
| 1768 | Berlin | Germany | 8.03 | 8.498 | 8.714 | 8.727 | | 6.79 | 8.012 |
| 1769 | Berlin | Germany | 8.46 | 8.408 | 8.656 | 8.725 | 8.626 | =AVERAGE(D2:D21) | 7.870 |
| 1770 | Berlin | Germany | 8.5 | 8.400 | 8.607 | 8.741 | 8.559 | 7.69 | 7.758 |
| 1771 | Berlin | Germany | 7.45 | 8.116 | 8.405 | 8.597 | 8.444 | 7.85 | 7.646 |

The same logic was applied to the global average temperatures.

### 3. Create a line chart comparing Berlin's temperature with global temperatures

After trying the line chart with all the calculated moving average, I decided to use the 15-years MA to do the analysis. The reason is because the noise is noticeably less than the 10- and 5-years MA but there's not much difference with the 20 years-MA. This way we can reduce the noise without losing much data.




### 4. Observations

- From the chart the average temperature in Berlin has always been slightly hotter than the global temperature. The difference between both temperatures is

similar over the years, the average difference between temperatures is 0.558 Celsius degrees.

- There's only one period between the years 1819 and 1825 that the difference is higher than usual. During this period the global average temperature had a noticeable decrease, but in the case of Berlin the average temperature maintained quite stable and even increased.
- In overall, aside of the previously mentioned period between the years 1819 and 1825, the global average temperature has been smoothly increasing. In the case of Berlin, it has also increased along the years, but it has more spikes. This can also be seen by calculating the percentage of difference in the temperature from year to year. In the case of Berlin, the highest decrease of temperature from year to year was of 28,746% between the years 1863 and 1864. And the highest increase was of 40.819% between the years 1871 and 1872. In the case of global temperatures, the difference is not so wide: the highest decrease of temperature from year to year was of 17,51% between the years 1767 and 1768. And the highest increase was of 13.42% between the years 1768 and 1769.
- Both temperatures averages, global and Berlin's, have been increasing over the years, something that has been specially accentuated since 1980. This strong relationship between the average temperatures can be validated by their correlation value of 0.897.

## 5. Additional observations

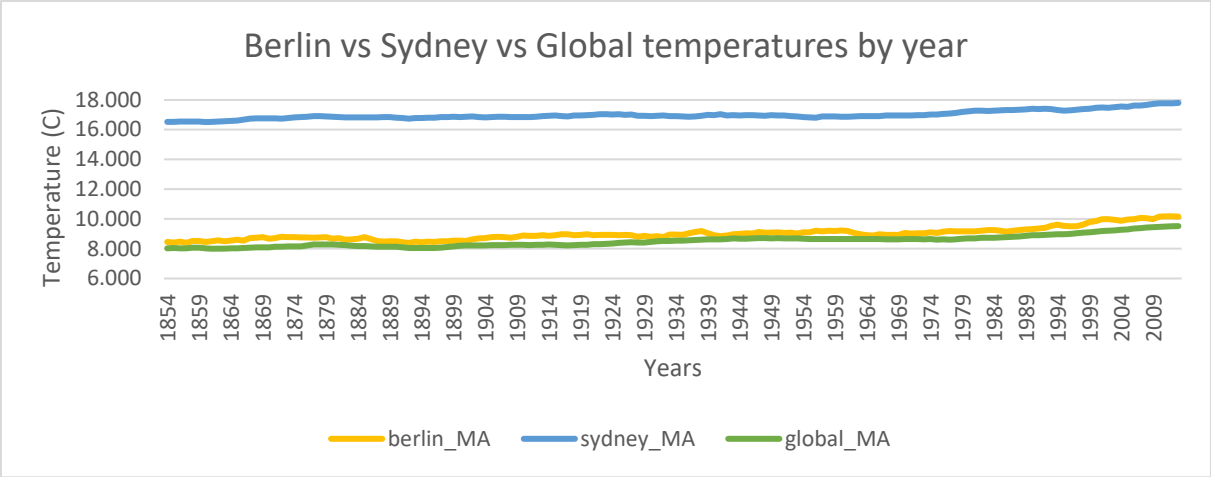
After having observed the comparison between Berlin and the global temperatures I decided to add to the analysis the data from Australia. My interest in Australia was because of the recent ecological events, and I wanted to see if over the years Australia had a bigger impact because of the global warming than Germany or not. I chose Sydney. Following the same logic than before I extracted the data with the following SQL query:

```
SELECT berlin.year
      , berlin.avg_temp as berlin_avg_temp
        , Sydney.avg_temp as Sydney_avg_temp
      , global.avg_temp as global_avg_temp
FROM city_data as berlin
INNER JOIN global_data as global
on berlin.year = global.year
INNER JOIN city_data as Sydney
on berlin.year = Sydney.year
WHERE berlin.country = 'Germany'
AND berlin.city = 'Berlin'
AND berlin.avg_temp IS NOT NULL
AND Sydney.country = 'Australia'
AND Sydney.city = 'Sydney';
```

Given that more data is missing from Sydney than from Berlin and the global data the comparison starts in the year 1854.

I downloaded the results in a CSV, opened them in Excel and calculated the Moving average with a gap of 15 years.

I generated the following graph:




The first observation is that, given that many years have been left out, to compare the data with Sidney, Berlin's and the global seem to have an even stronger relationship. Mostly affected by excluding the time period between the years 1819 and 1825, where was observed the biggest difference among the temperatures before.

Unlike what I was expecting, the average temperature in Sidney has increased over the years like the other two measures. In fact, the correlation coefficients among the temperatures is quite similar:

| | *Berlin_MA* | *Sydney_MA* | *Global_MA* |
|---|---|---|---|
| Berlin_MA | 1 | | |
| Sydney_MA | 0.917644729 | 1 | |
| Global_MA | 0.956859194 | 0.90924885 | 1 |

In these calculations can also be seen that the correlation coefficient between Berlin and the global temperature shows the even stronger relationship that was observed above.