Udacity: Data Analyst – Project 1

Exploring Weather Trends

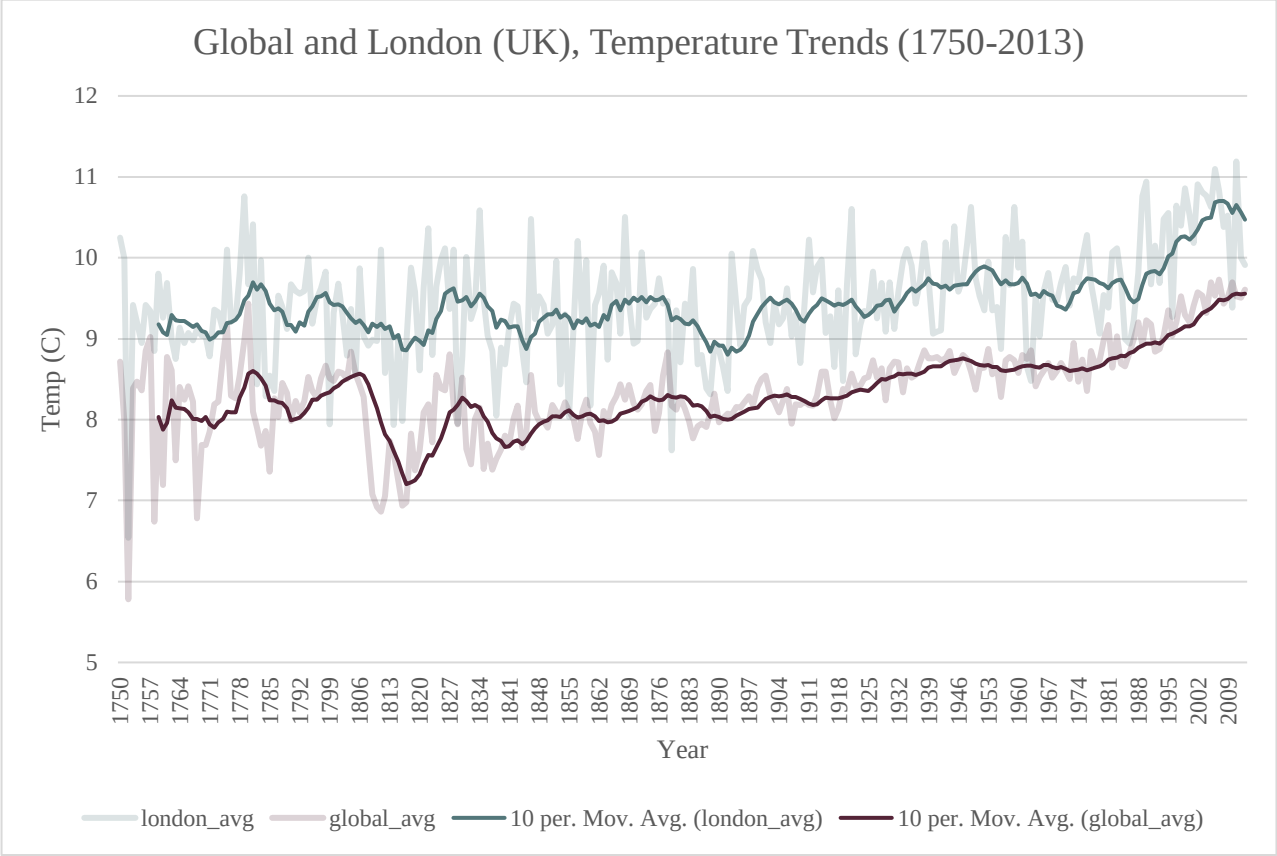
Gurpreet S Rai

Project outline:

To analyse both the global average and city of London (UK) temperature datasets, in the period from 1750 to 2013, and to compare temperature trends across this period.

- To extract the global and London temperature data from a database, with an SQL query, and export to CSV.
- To open, and create a visualisation of the data within Excel, in the form of a line chart. Plotting *moving averages*, to make tends more observable.
- To make at least four observation about the similarities, differences, and overall trends, between the global average and city of London temperature data.
- To outline all of the steps taken to prepare the data to be visualised in the chart.

Moving Averages Line Chart:



Observations:

From the Global and London Temperature Trends (1750-2013) Line chart above, there are obvious and clear similarities and differences between the global and London weather data trends.

The similarities include, the general alignment of the year to year data points, which is most clear in the moving average (opaque) plots. They almost perfectly follow the same trajectory on the 10year moving average plots, which shows a clear correlation between the city of London and the global average temperature. London being aligned with the global average temperature might seem an expected result, but it does show that any local regional temperature variance is low compared with the global average.

The main differences between the global and London datasets, is the larger temperature variance London experiences year-on-year, than the global average (though still visible on the moving average plot). This would be expected of course if you consider the global dataset is likely to have year-on-year variances smoothed out by the numerous locations points the data is pooled from.

The London temperature dataset overall, shows a consistent 1C (approx.) higher temperature than the global average. This is most evident on the moving average plot, and is consistent over the last several hundred years. See below *Further observations*.

There is a notable dip in the global average temperatures between the years 1808-1820. The London data also shows a drop in temperature during this period, though not as pronounced as the global average. Further research has shown this was due to two large volcanic eruptions, the 1808/1809 mystery eruption[1] and the 1815 eruption of Mount Tambora[2].

Overall, the data for both the city of London and global average, show a general increase in temperature over the years. This trend starts to become obvious from the beginning of the twentieth century, with a notable rate of increase in temperature of both London and the global average from the 1980's onwards, where the warming trend is starting to show a precipitous increase. This warming trend is most likely due from human-produced greenhouse gasses into the atmosphere[3] which have been increasing from the beginning of the industrial revolution[4].

Outline of steps - Data to Visualisation:

Accessing Data with SQL:

The Data Schema was determined by drilling through each of the three tables (city_day, city_list, global_data).

Through SQL query (select * from city_list where city = 'London') – I was made aware of a second city of London in Canada. Therefore, requiring a further restriction of 'Country' to my final SQL query when selecting my city data.

I decided the following datasets where required, to provide the information needed for both formula and plots: 1) Average London (UK) temp data, 2) Average global temp data, 3) Year data was collected.

To obtain all three datasets, I wrote and executed the following SQL query:

```
1  select global_data.year as Year, city_data.avg_temp
   as London_Avg, global_data.avg_temp as Global_Avg
2  from city_data, global_data
3  where city_data.year = global_data.year
4  and city_data.city = 'London' and city_data.country
   = 'United Kingdom'
5  order by global_data.year
```

SQL query methodology

- Select required fields - year, avg_temp (global_data), avg_temp (city_data)
- From two tables – city_data, global_data.
- On the join condition of 'year' from both tables.
- Applying restrictions of City and Country on the city_data.
- Ordering data output by 'year'.

Producing Moving Averages from CSV Data:

The output from the above query was downloaded as a CSV, and opened in Excel.

A standard line chart was created from the CSV Data, and by adding a Trend Line to both plots, produced a 10year Moving Average plot. I decided to keep the original plots as reference, but reduced the transparency to help highlight the Moving Averages plots.

Further observations:

Correlation Coefficient – Global Average vs London (UK) average:

Using the Excel CORREL Function:

$$Correl(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

The Correlation Coefficient between the two datasets was found to be **0.563099.**

This indicates a medium to strong correlation between the global average and London average temperature datasets.

Average Temperature Difference Between Datasets:

London was found to have an average temperature difference of **+1.1C**, relative to the global average.

This average temperature difference was calculated by taking the averages for the global average and London average, and finding the difference between the two.

Citations:

[1] 1808/1809 mystery eruption – Wikipedia (https://en.wikipedia.org/wiki/1808/1809_mystery_eruption)

[2] 1815 eruption of Mount Tambora – Wikipedia (https://en.wikipedia.org/wiki/1815_eruption_of_Mount_Tambora)

[3] Climate change causes - NASA (https://climate.nasa.gov/causes/)

[4] Greenhouse gas changes since the Industrial Revolution - ACS (https://www.acs.org/content/acs/en/climatescience/greenhousegases/industrialrevolution.html)