Patrick Spears

Data Analyst Nanodegree

11/21/2017

Explore Weather Trends

Extract

To retrieve the data from the database, I iterated through several versions of a query. On my first attempt at pulling the data, I noticed that there were several nulls -- ~3.5% of the records were missing a value for a given year.

```
SELECT count(*)
FROM city_data
WHERE avg_temp Is Null;
```

Having discovered that, I wanted to choose a city that had one of the more continuous sets of data. I compared the count of Nulls for each city in the city_data set:

```
SELECT city, count(*)
FROM city_data
WHERE avg_temp Is Null
GROUP BY city
ORDER BY count(*) asc;
```

Fortunately, Chicago had only four null values, so it was an ideal choice for the project exploration. I decided to settle on Chicago as my primary city of choice, but to also pull a large set of data for other cities to show consistency across any trends. Since we're interested in using a moving average for this project, I used a window function to calculate the moving average for each city and extract the global averages in a single dataset.

```
SELECT year,
       city,
       avg_temp,
       AVG(avg_temp)
            OVER(PARTITION BY city
                  ORDER BY year ROWS BETWEEN
                  UNBOUNDED PRECEDING AND CURRENT ROW)
            as moving_average_temp
FROM city_data T1
WHERE city IN ('Chicago', 'Sydney', 'Copenhagen', 'Munich', 'Shenzhen',
                  'Manila', 'Moscow', 'Stockholm', 'Johannesburg')
      AND year > 1749
UNION SELECT year,
       'Global',
       avg_temp,
       AVG(avg_temp)
            OVER(ORDER BY year ROWS BETWEEN
            UNBOUNDED PRECEDING AND CURRENT ROW)
FROM global_data
WHERE year > 1749
ORDER BY city,
         year;
```

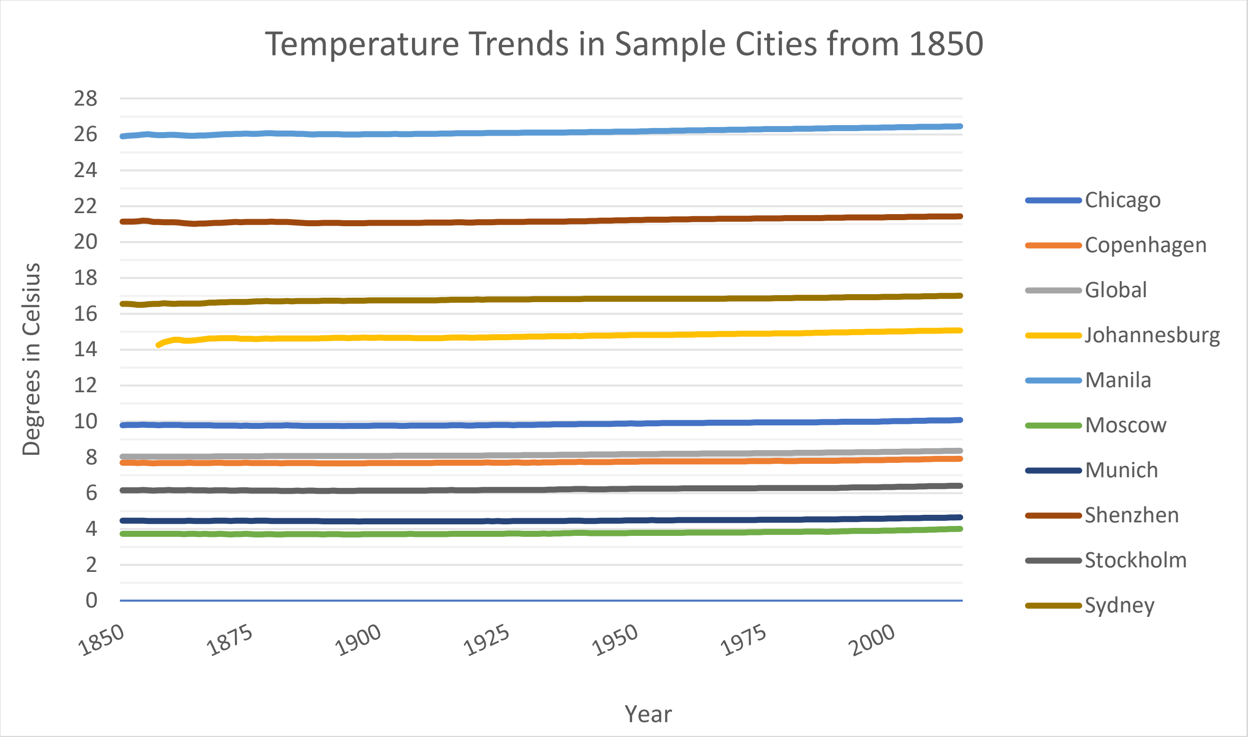
Transform

The dataset I extracted is structured for ease of extracting through SQL, but it's not strictly formatted that Excel can easily create graphs through it. I restructured the data using an index-match array function to resemble a dataframe-like structure: an array of years, followed by an array for each of the cities I selected.

One important thing to note about the data is that many of the cities did not have a full range of data; either collection started later for these cities or there were otherwise gaps in the data. For example, the data for Chicago starts in 1743, but is missing data for 1746-1749. The global average temperatures start in 1750, it seemed like a natural choice to begin my dataset at 1750. However, the comparator cities included started at various other years. Consequently, the analysis of multiple cities begins at 1850, the year a plurality of entities had consistent data.

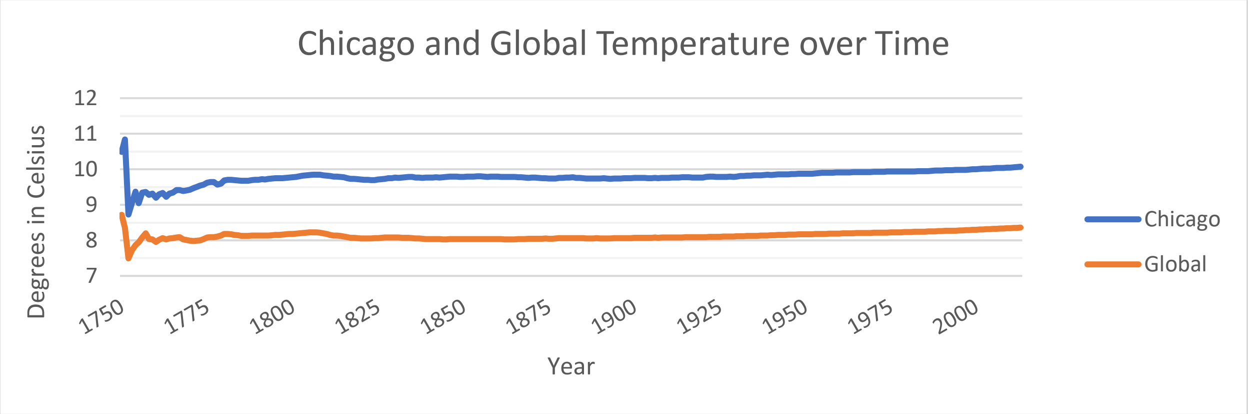
Analyze

I plotted the cities' moving average temperature by year in line graphs. The first chart indicates, for comparison, that each city selected shows a similar general trend of increase temperature over time. To me, the trend seems more extreme in the warmer cities, such as Manila, in comparison with extremely cold cities like Moscow.



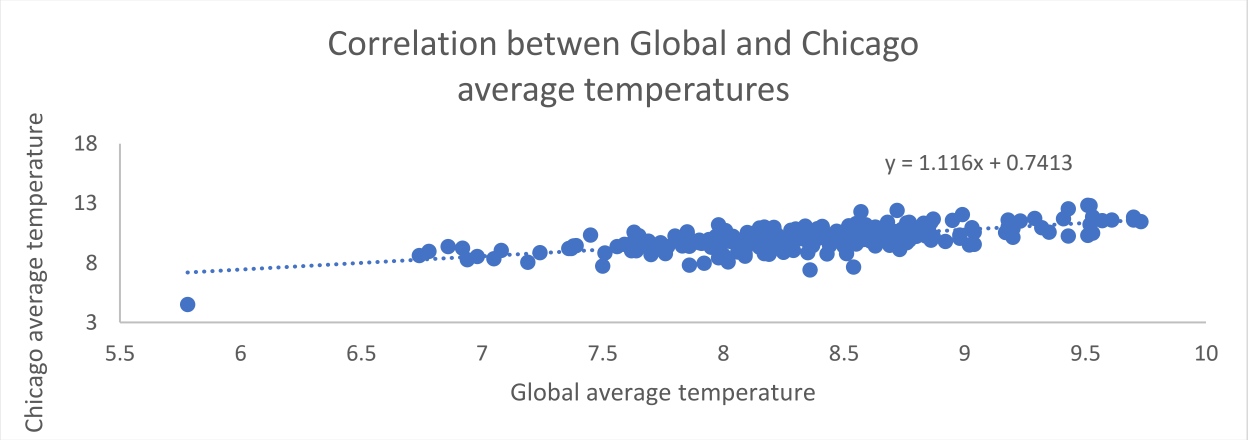
When we review the Chicago and Global temperatures in closer detail, we can see the general trend over a greater period of time, because the analysis starts in 1750. There are two important things to note off the bat in the plot:

1. 1750 appears to be an abnormally warm year, both in Chicago and globally. There is further a sudden dip immediately afterward. These spikes could represent outliers in the data, or inconsistencies with recording it.
2. Generally, the current trend of steadily increasing temperature begins around 1850 – as noted in many analyses of the weather, this coincides with the height of the industrial revolution. There is a more dramatic increase around 1925.

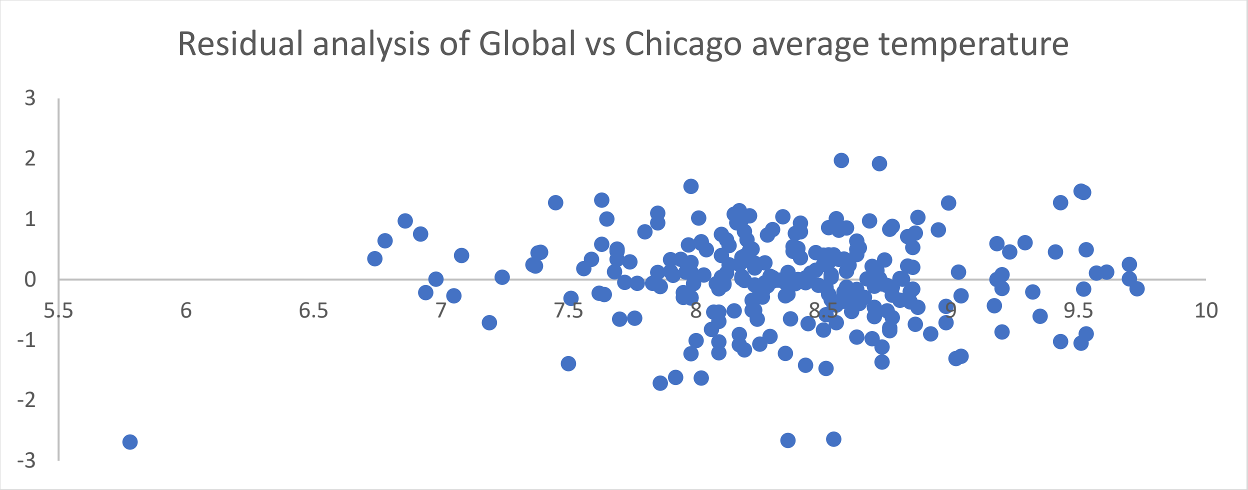


A widely disseminated fact today, the plot above indicates the general increase of temperature to be around **1.5° C**.

From the graph, we can see that Chicago is generally hotter than the global average temperature and, over time, follows a similar trend. The trend is so similar, in fact, that it warrants further analysis. To dive deeper, I examined the raw average temperatures globally and within Chicago.



We can see in this scatterplot that there is a positive correlation between the temperatures; the Pearson coefficient is 0.7785, indicating a strong relationship between the two. The predictive equation – based on the line of best fit – can be used to predict Chicago's temperature from a given global average temperature value. To verify that linear regression is a good fit for this data set, I plotted the residual values and found a random pattern. This means that the regression model is appropriate for the dataset.



However, for the average temperatures, the r^2 is only 0.4477; consequently, the variance between the model doesn’t well explain the variability in the data set. This suggests that a further study in variability might be productive.