Explore Weather Data

To analyze the weather data, I take four steps. I list each of the steps and the tools I used for it in the following paragraphs.

1. Extract the data

In the first step, I used SQL to extract weather data in the original tables. The first query is used to extract the temperature data of the nearest big city, Chicago. The second query is used to extract the global weather temperature.

When I extracted data, I found that some of the values in the avg_temp column in the city_data table were missing. Thus, I only chose the values that were not null.

Query 1:

```
select year, city, avg_temp
from city_data
where city = 'Chicago' and avg_temp is not null
;
```

Query 2:

```
select *
from global_data
where avg_temp is not null
;
```

2. Open up the CSV

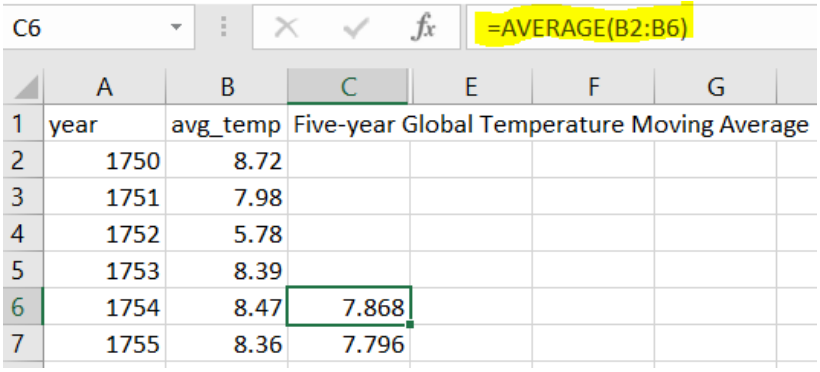
I was required to open up the CSV files downloaded from the Udacity website in the second step. I used Microsoft Excel to complete the requirement.

3. Calculate the moving average and create a line chart

After opening the files in Excel, I wanted to use the AVERAGE() function to calculate the moving average of the temperature data. Before calculation, the first thing I did was reviewing and cleaning the data. I found that the Chicago weather data were from 1743-1745 and 1750-2013. (Data for 1746-1749 were deleted because they were null values.) The global weather data, however, was from 1750-2015. To make the comparison more meaningful, I chose the data from 1750-2013 for both tables and ignored the other data.

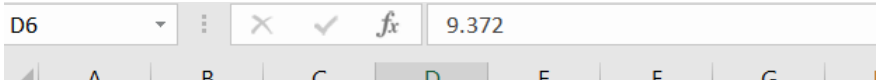
The second thing I did was determining the length of the moving average interval. This was also one of my key considerations when I visualized the data. At first, I tried to calculate the three-year moving average, but I found the chart was very volatile – there were so many small changes in temperatures. Then, I tried the ten-year moving average, but the chart became too smooth, omitting essential details that would impact the analysis. Thus, I chose to use the five-year moving average to create the line chart.

To calculate the five-year temperature moving average of Chicago and global from 1750 to 2013, I used the AVERAGE() function. Take the global temperature data as an example, I created the third column named "Five-year Global Temperature Moving Average". Then, in this column, I used the AVERAGE() function to calculate the averaged temperature of the first five years, namely 1750-1754. You can refer to the picture below to understand how I completed this process.



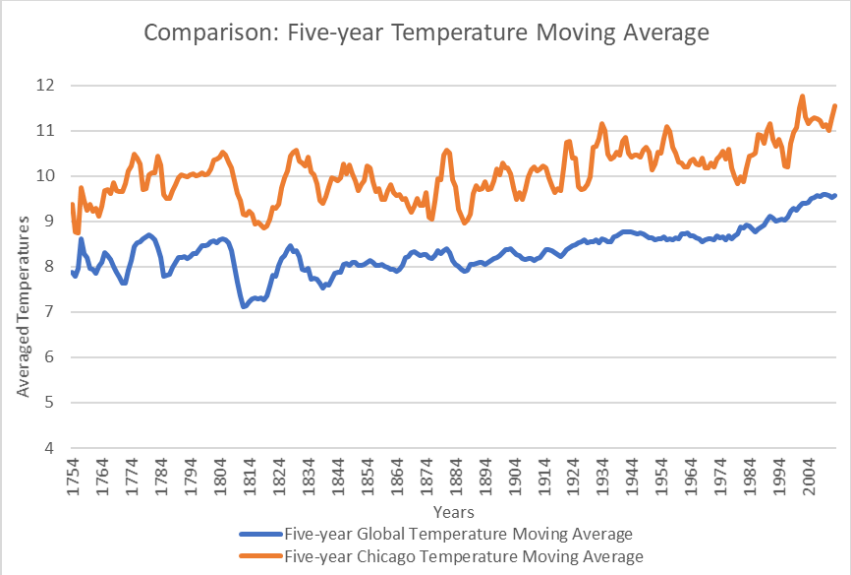
	A	B	C	E	F	G
1	year	avg_temp	Five-year Global Temperature Moving Average			
2	1750	8.72				
3	1751	7.98				
4	1752	5.78				
5	1753	8.39				
6	1754	8.47	7.868			
7	1755	8.36	7.796			

Then, I dragged down the formula to get the five-year moving average for each value in the global temperature data. I repeated these two steps to calculate the five-year moving average for the Chicago temperature data. After calculation, I pasted the moving average value for Chicago temperature to the global temperature spreadsheet to create a line chart containing the two types of data.



	A	B	C	D	E	F	G
1	year	avg_temp	Five-year (	Five-year Chicago Temperature Moving Average			
2	1750	8.72					
3	1751	7.98					
4	1752	5.78					
5	1753	8.39					
6	1754	8.47	7.868	9.372			
7	1755	8.36	7.796	8.756			

Besides the length of the moving average intervals, I also considered making the chart more convincing and understandable. Hence, I added the chart title, axis titles, and legends. I used different colors for different lines. Moreover, I adjusted the y-axis to place the lines in the center of the graph, showing the trend and the details more clearly. You can see the final version of my chart below.



4. Observations

As is shown above, my city is hotter on average than the global temperature. The difference is consistent over time because the average temperatures in Chicago are always higher than those around the globe.

The temperature changes in my city are generally the same as the changes around the globe. However, the temperature changes in Chicago are more volatile than the changes in global temperature.

The overall trend is that global temperatures are rising in the past two hundred years. The world is getting hotter than it was in the 1700s. The trend has not been so consistent over the past few hundred years. During the late 1700s, early- and mid-1800s, the average global temperatures fluctuated violently and decreased about 1 degree Celsius three times.

Similar to the global weather, the trend of the Chicago temperature is also not consistent. Although the weather in Chicago is getting hotter overall, it fluctuates frequently. Michigan Lake might impact the fluctuations.

5. Using R to build a linear regression model

(Notice: I cleaned the data in the previous steps, so I only used the Chicago and global temperature data from 1750 to 2013. Including other data in the dataset will result in different models and values.)

To use the global temperature to estimate the temperature in Chicago, we can build a linear regression model. I used lm() function in R to achieve this goal. The summary of the model is listed below.

```
Call:
lm(formula = avg_temp_chicago ~ ., data = Book1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.69153 -0.34242  0.01572  0.44784  1.97493

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.74128    0.64158   1.155    0.249
avg_temp_global 1.11596    0.07657  14.575   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7142 on 262 degrees of freedom
Multiple R-squared:  0.4477,    Adjusted R-squared:  0.4456
F-statistic: 212.4 on 1 and 262 DF,  p-value: < 2.2e-16
```

We can see that the t statistic and the F statistic are both lower than the default alpha, 0.05, which means that the predictor, avg_temp_global, is statistically significant. The adjusted R squared is 0.4456. It means that the independent variable can explain 44.56% variability in the dependent variable.

I also used the cor() function to check the correlation coefficient. The result is listed below. The correlation coefficient is 0.6691357.

```
                 avg_temp_chicago avg_temp_global
avg_temp_chicago        1.0000000       0.6691357
avg_temp_global         0.6691357       1.0000000
```