Exploring Weather Trends

Introduction

In this project, I am going to analyze local and global temperature data. I will compare the temperature data of the city where I live to the global temperature data, then answering the following questions: Is my city hotter or cooler than the average global temperature? Does my city temperature have the same trend as the global temperature? Finally, is the global temperature increasing, decreasing, or staying the same over the years?

The main work will be divided into two phases, data wrangling, and data analysis.

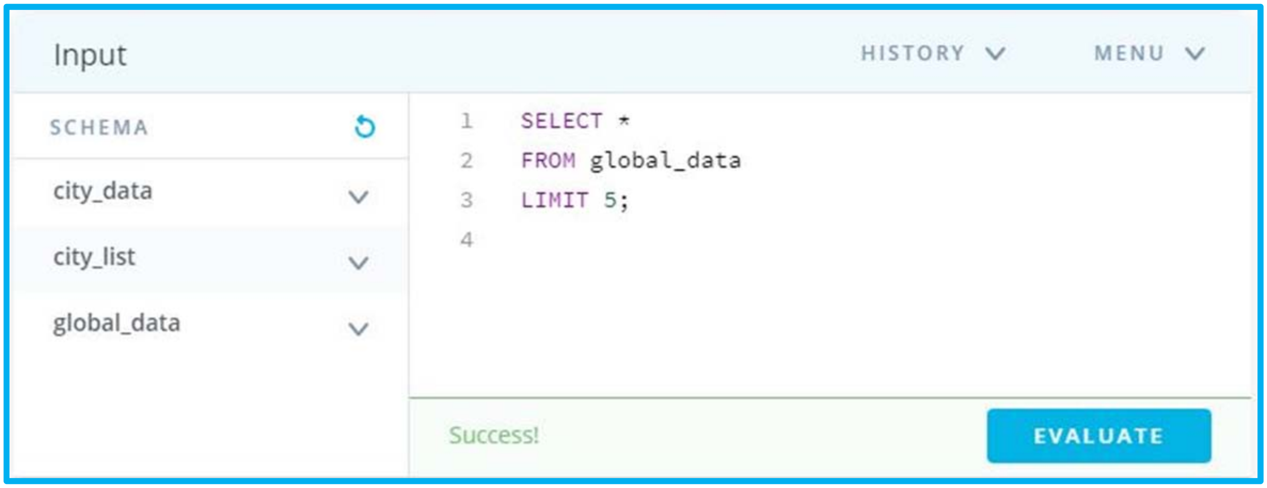
In data wrangling phase, I will do simple wrangling, such as filtering, querying, and collecting data using SQL. I will also include every input code and output result. The **input code** will have a **blue** border, while the **output result** will have a **green** border.

In the data analysis phase, I will use Microsoft Excel to analyze the data and make visualization.

Data Wrangling

In this phase, I will use SQL to do simple data wrangling. First, let us get a sense of the data by viewing the first five rows of each file: **global_data**, and **city_data:**

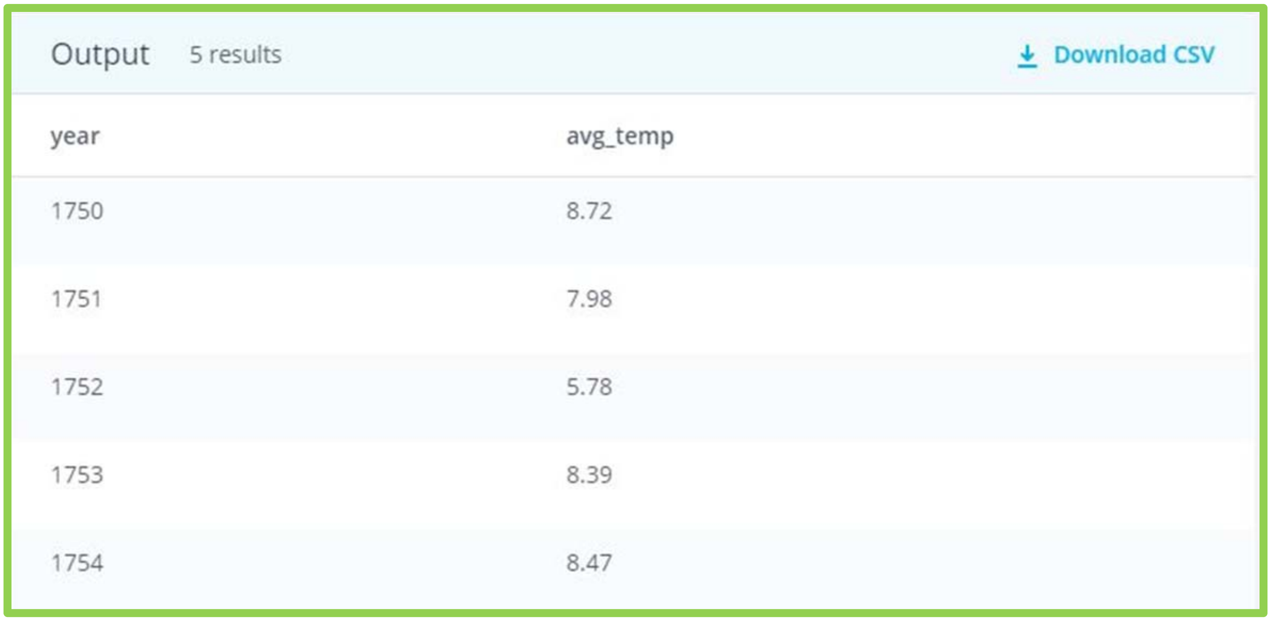
Here is the input code for the first file, **global_data:**



```
SELECT *
FROM global_data
LIMIT 5;
```

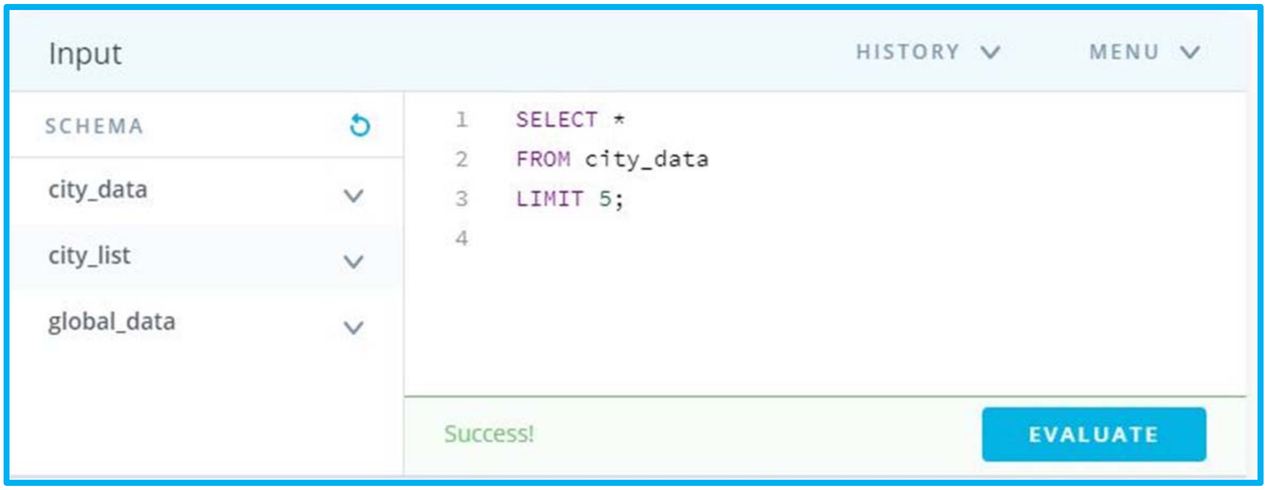
Success!

The output:



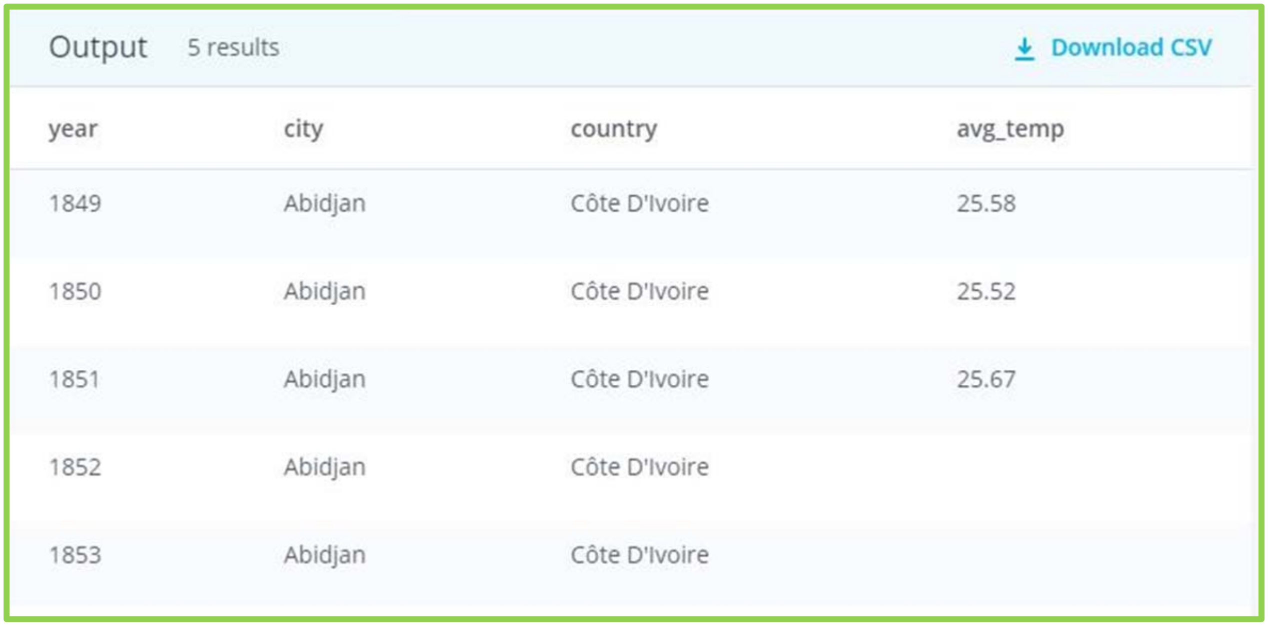
year	avg_temp
1750	8.72
1751	7.98
1752	5.78
1753	8.39
1754	8.47

And the input code for the second file, **city_data**:



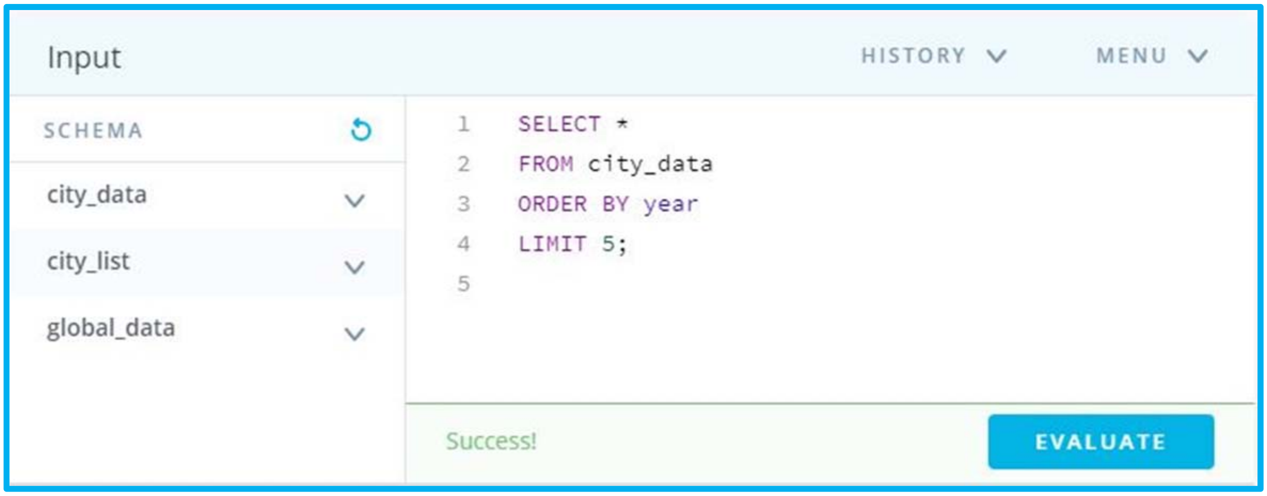
```
SELECT *
FROM city_data
LIMIT 5;
```

The output:



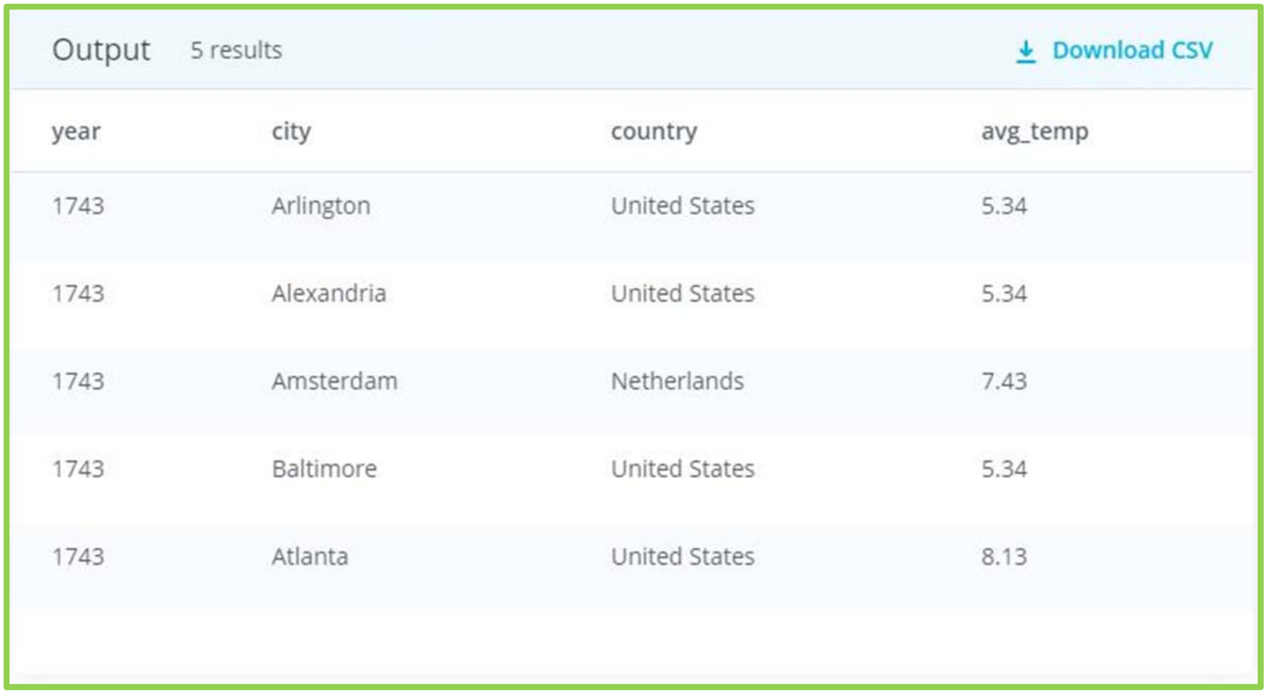
year	city	country	avg_temp
1849	Abidjan	Côte D'Ivoire	25.58
1850	Abidjan	Côte D'Ivoire	25.52
1851	Abidjan	Côte D'Ivoire	25.67
1852	Abidjan	Côte D'Ivoire	
1853	Abidjan	Côte D'Ivoire	

As we can see from both output tables, the **global_data** starts at the year 1750 while the **city_data** starts at the year 1849, so I am going to sort data in **city_data** by year to see what year it starts with.



```sql
SELECT *
FROM city_data
ORDER BY year
LIMIT 5;
```

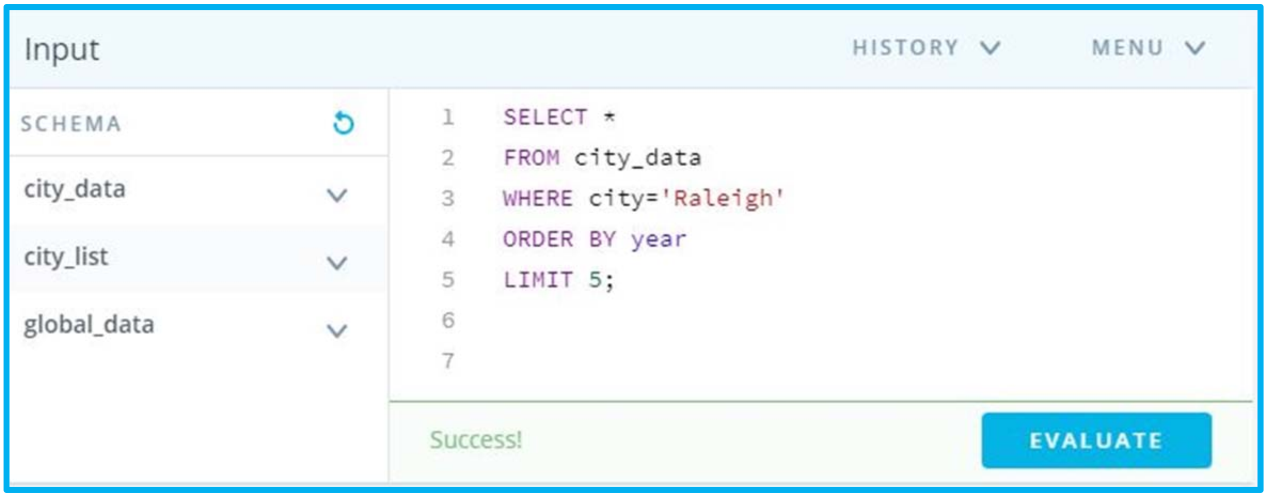
Results:



year	city	country	avg_temp
1743	Arlington	United States	5.34
1743	Alexandria	United States	5.34
1743	Amsterdam	Netherlands	7.43
1743	Baltimore	United States	5.34
1743	Atlanta	United States	8.13

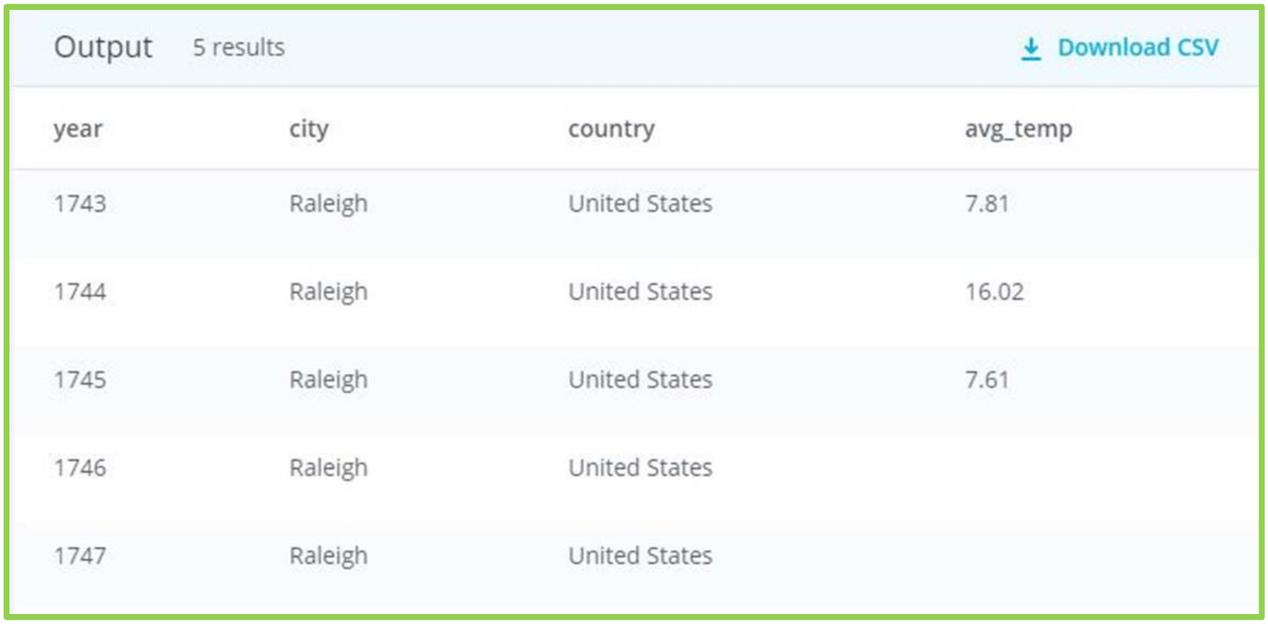
From the output above, we see the **city_data** starts with 1743. There still a difference in years between the two tables that I need to address later.

Now, I am going to filter the **city_data** to get information about the city where I live, which is Raleigh as follows:



```
SELECT *
FROM city_data
WHERE city='Raleigh'
ORDER BY year
LIMIT 5;
```

Success!



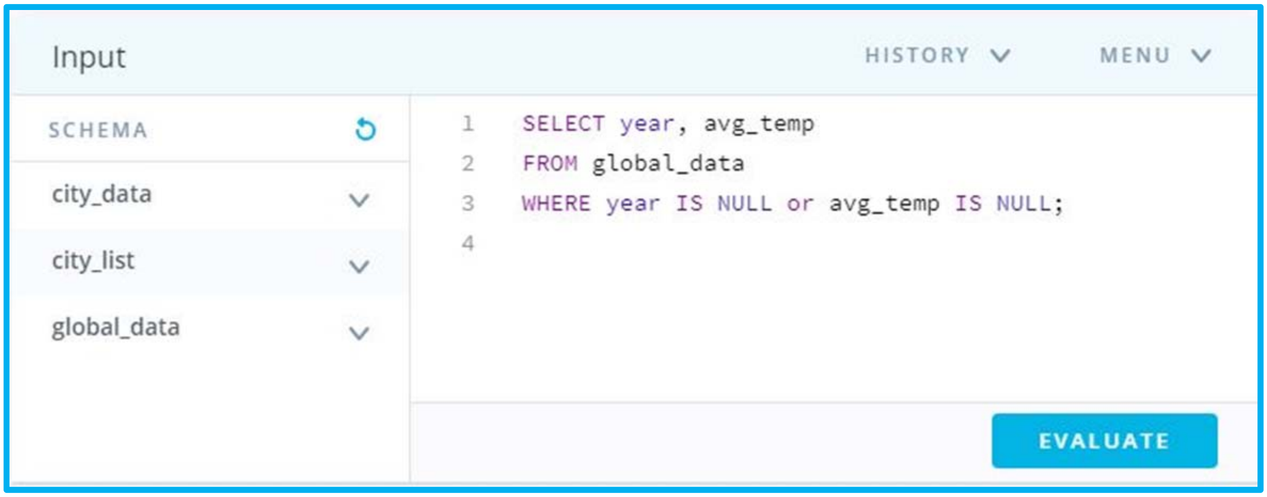
year	city	country	avg_temp
1743	Raleigh	United States	7.81
1744	Raleigh	United States	16.02
1745	Raleigh	United States	7.61
1746	Raleigh	United States	
1747	Raleigh	United States	

From the output above, we see the year where the data were collected for the city of Raleigh is also 1743, while the first year of **global_data** is 1750. Therefore, we have seven years of missing data in the **global_data.**

We can also notice there are some missing values in avg_temp field in the **city_data**.

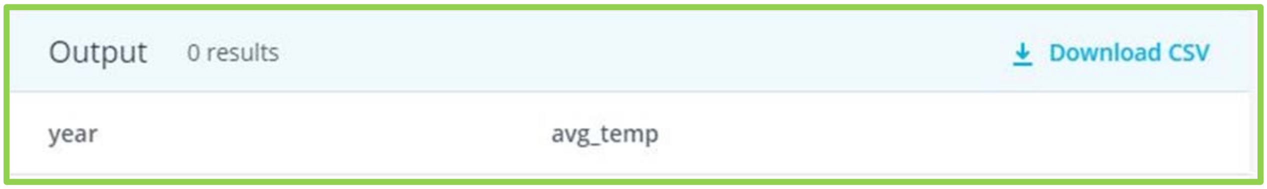
I am going to look for missing values for every field in both the **global_data** and **city_data**.

I will start with **global_data.**



```sql
SELECT year, avg_temp
FROM global_data
WHERE year IS NULL or avg_temp IS NULL;
```

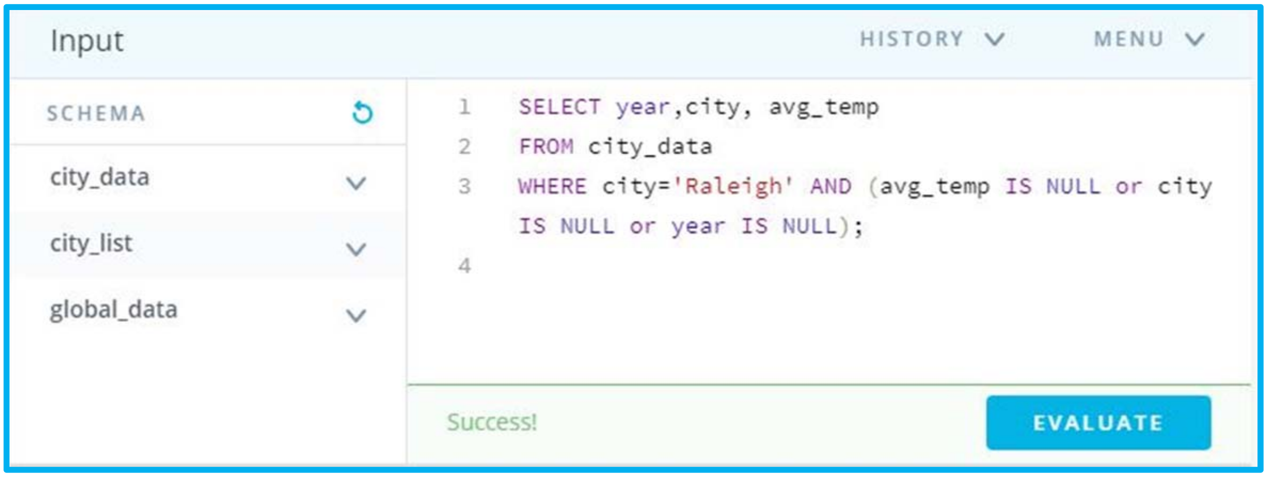
The result:



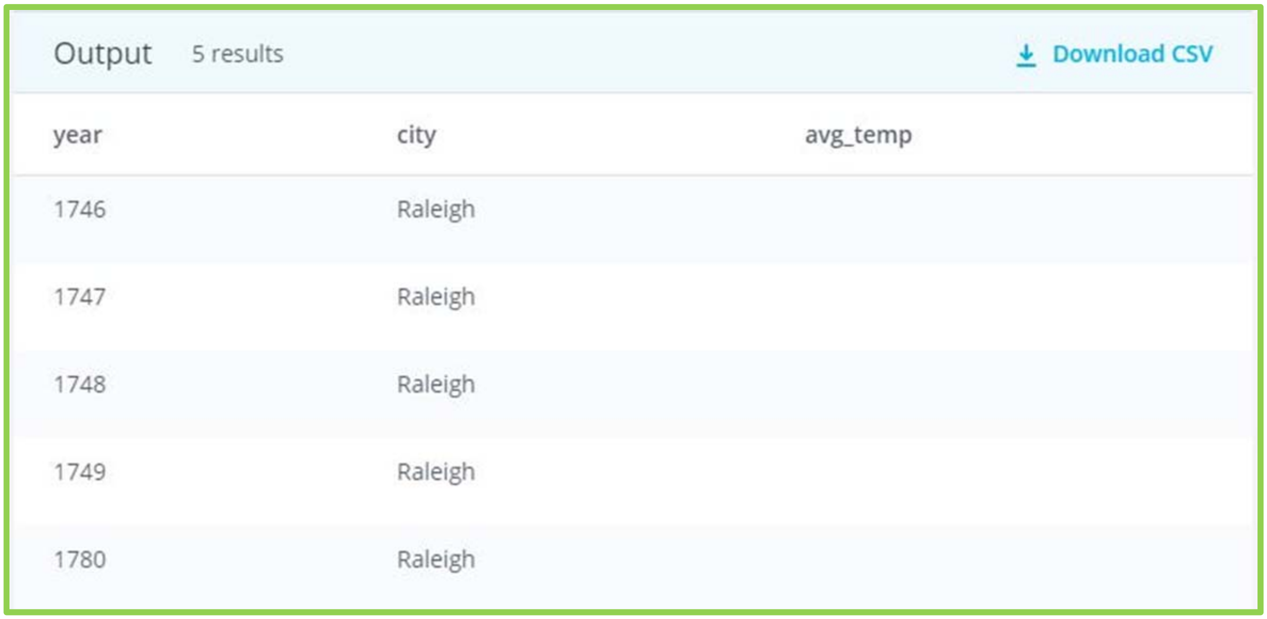
year	avg_temp

We see the output has 0 results, which means we don’t have any missing values in the **global_data.**

Now, Let us check for the missing values in the **city_data** for the city of Raleigh



```sql
SELECT year,city, avg_temp
FROM city_data
WHERE city='Raleigh' AND (avg_temp IS NULL or city IS NULL or year IS NULL);
```



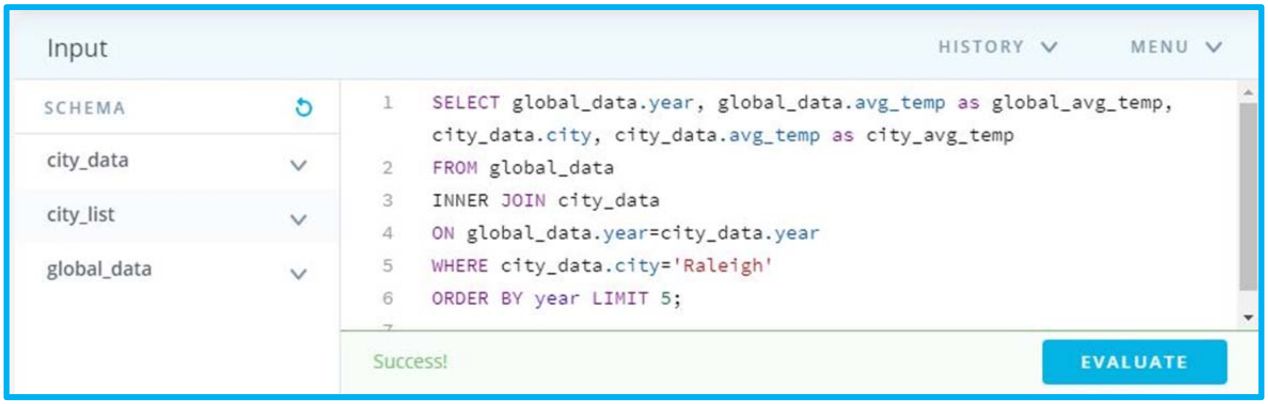
year	city	avg_temp
1746	Raleigh	
1747	Raleigh	
1748	Raleigh	
1749	Raleigh	
1780	Raleigh	

From the output above, we have five missing values in avg_temp for the years 1746, 1747, 1748, 1749, and 1780.

There are many ways to deal with missing values. In this project, I will discard the entries with missing value.

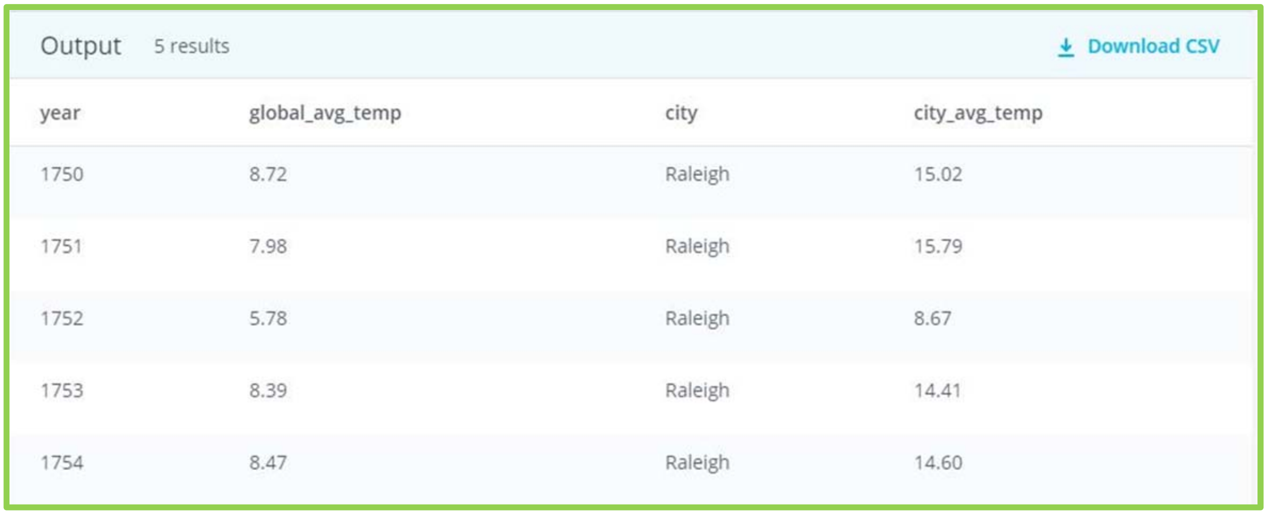
Joining Tables

To remove the entries of the **city_data** where years are before 1750, and match only similar years for both tables, I will inner join the **global_data** and the **city_data** tables on the year, which is the primary key.



```
SELECT global_data.year, global_data.avg_temp as global_avg_temp, city_data.city, city_data.avg_temp as city_avg_temp
FROM global_data
INNER JOIN city_data
ON global_data.year=city_data.year
WHERE city_data.city='Raleigh'
ORDER BY year LIMIT 5;
```

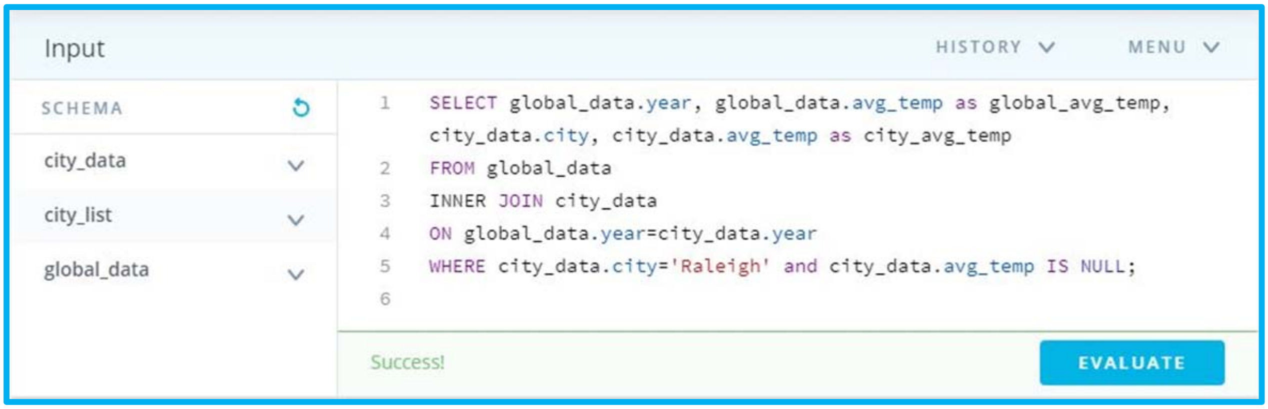
Success!



year	global_avg_temp	city	city_avg_temp
1750	8.72	Raleigh	15.02
1751	7.98	Raleigh	15.79
1752	5.78	Raleigh	8.67
1753	8.39	Raleigh	14.41
1754	8.47	Raleigh	14.60

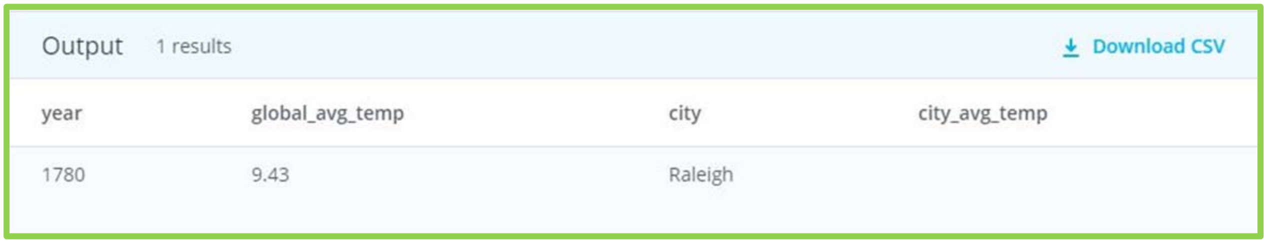
We can see the two tables were nicely joined, and now we have only one table that includes the global temperature, the city temperature for our specified city and the year where both average temperatures were collected.

Now, let us look for the missing values in the city_avg_temp. We know we have missing values for the following years: 1746, 1747, 1748, 1749, and 1780. Luckily, the years, 1746, 1747, 1748, and 1749 were removed by joining both tables. Therefore, I expect to have one missing value for the year 1780. Let us figure it out!



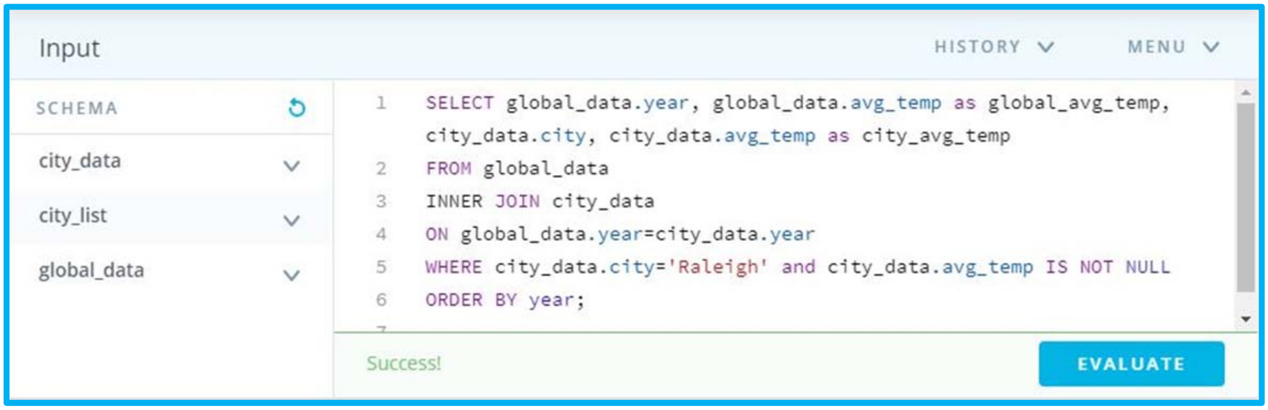
```
SELECT global_data.year, global_data.avg_temp as global_avg_temp,
city_data.city, city_data.avg_temp as city_avg_temp
FROM global_data
INNER JOIN city_data
ON global_data.year=city_data.year
WHERE city_data.city='Raleigh' and city_data.avg_temp IS NULL;
```

Success!



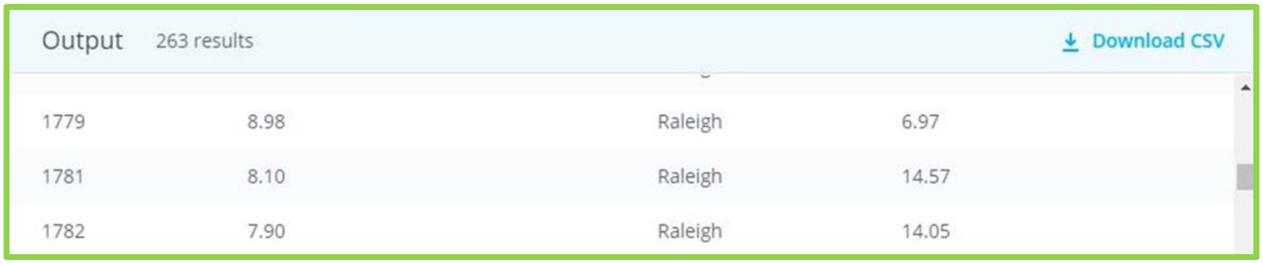
year	global_avg_temp	city	city_avg_temp
1780	9.43	Raleigh	

As expected, the only missing value is for the year 1780. Now, I am going to re-run the query to exclude that missing values



```sql
SELECT global_data.year, global_data.avg_temp as global_avg_temp, city_data.city, city_data.avg_temp as city_avg_temp
FROM global_data
INNER JOIN city_data
ON global_data.year=city_data.year
WHERE city_data.city='Raleigh' and city_data.avg_temp IS NOT NULL
ORDER BY year;
```

Since the output is too large, I included only part of it to show the excluded year as follows:



1779	8.98	Raleigh	6.97
1781	8.10	Raleigh	14.57
1782	7.90	Raleigh	14.05

So we see the year 1781 comes right after 1779 which means the year 1780 was removed.

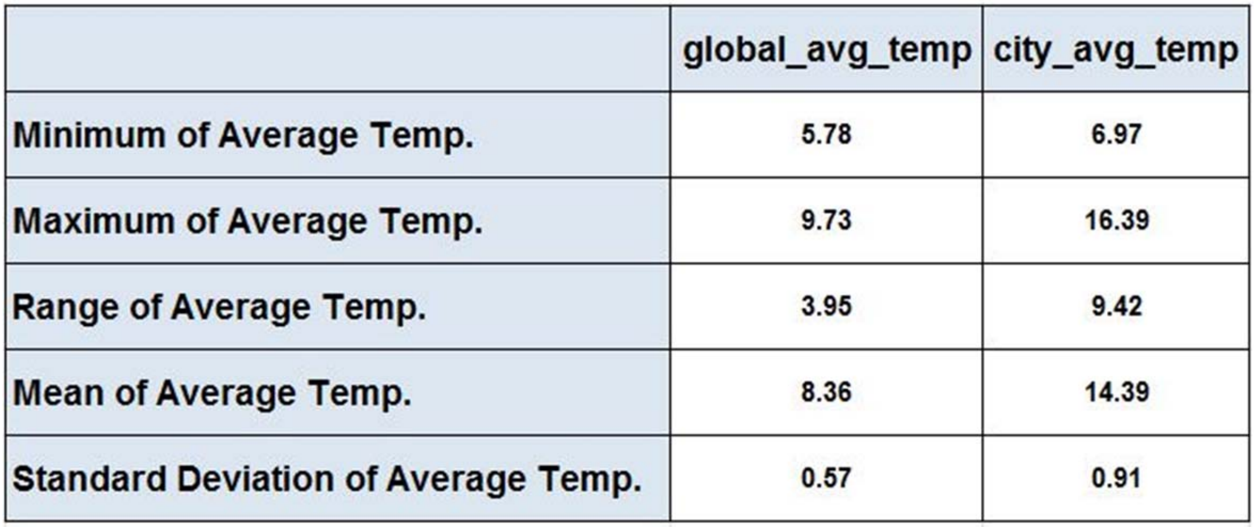
Now, I am going to save the resulting table in new file and use it for upcoming analysis.

To summarize, by now, we should have a file with average global temperature and average city temperature for the city of Raleigh from the year 1750 to 2013 excluding the year 1780.

Data Analysis

Descriptive Statistics

Before doing any analysis, I like to run some descriptive statistics to get an idea about the data in hand. Using Microsoft Excel, I took the following statistics:



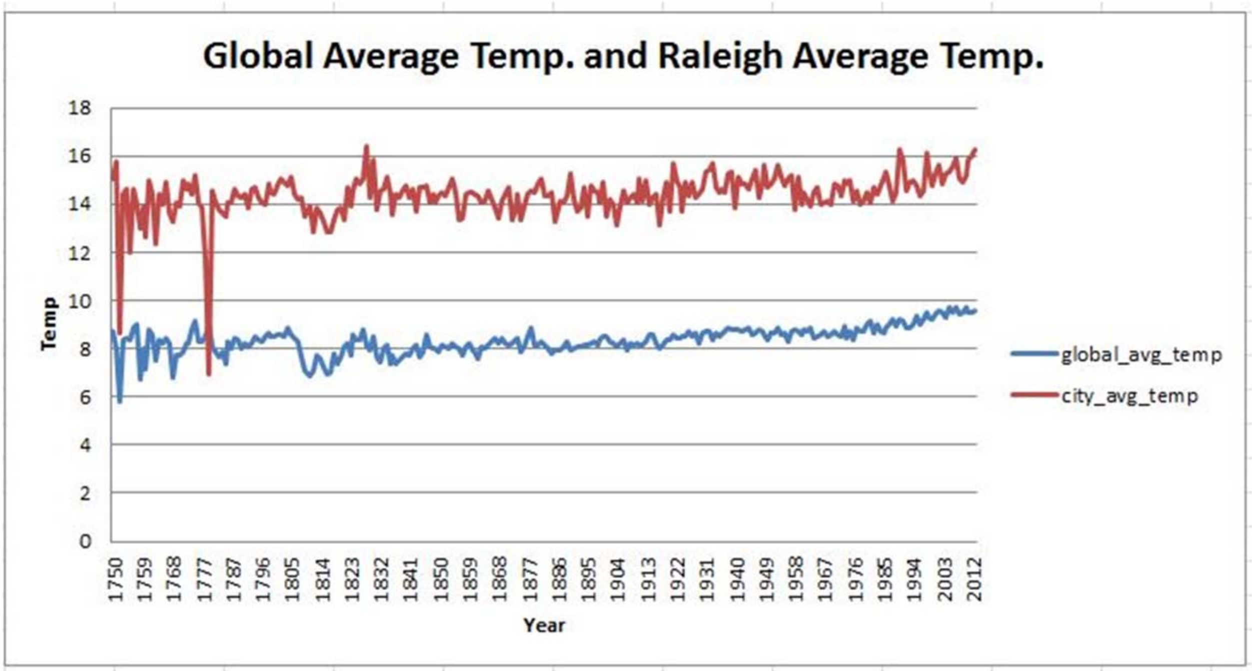
	global_avg_temp	city_avg_temp
Minimum of Average Temp.	5.78	6.97
Maximum of Average Temp.	9.73	16.39
Range of Average Temp.	3.95	9.42
Mean of Average Temp.	8.36	14.39
Standard Deviation of Average Temp.	0.57	0.91

From the above table, we notice couple of things:

- The mean of the city average temperature is higher than the mean of the global average temperature.
- From the range and standard deviation, we see that the temperature in the **city_data** fluctuate more than those in the **global_data**.

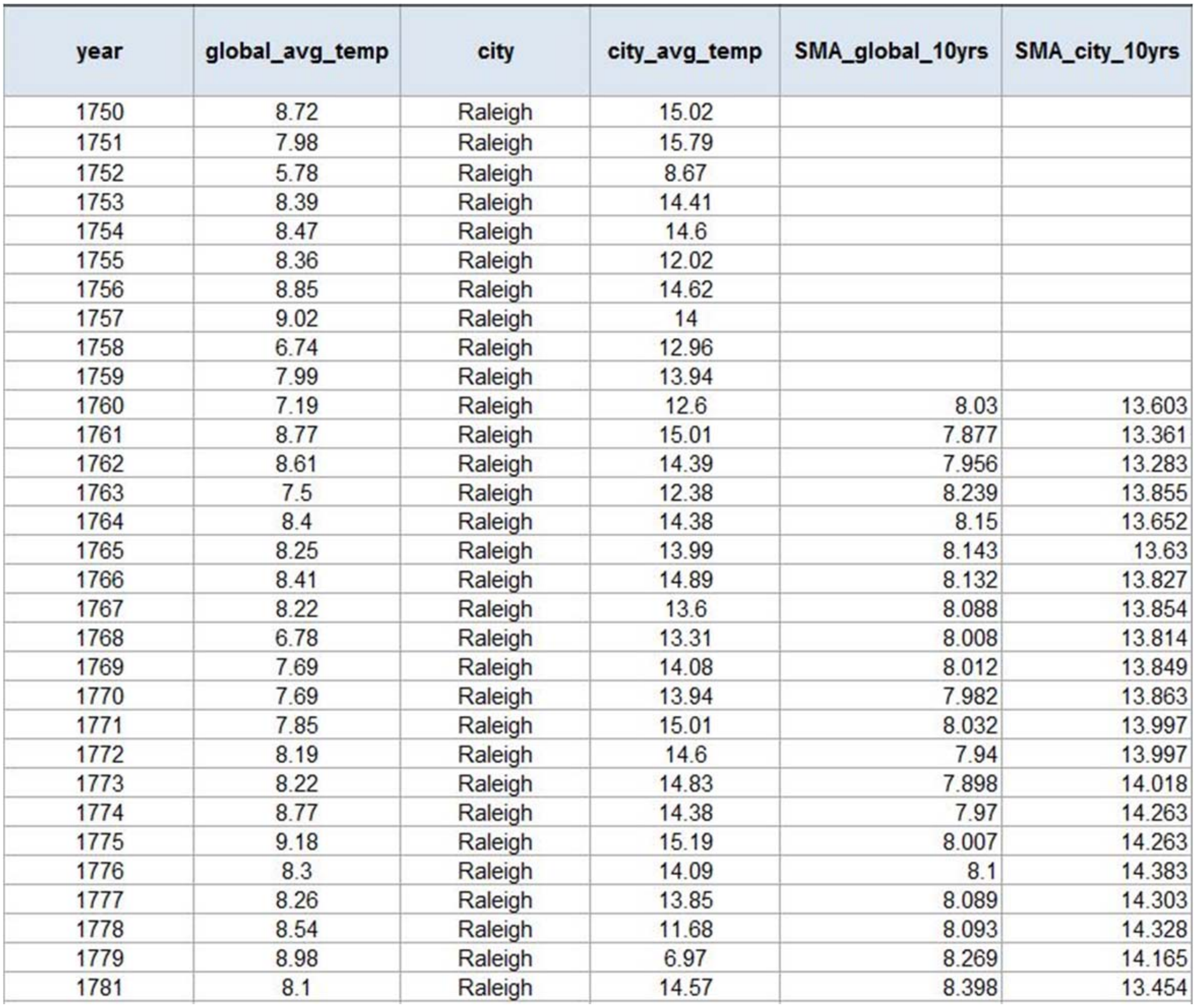
Line Charts

Here, I am going to graph a line chart of the original data for both global_avg_temp and city_avg_temp.



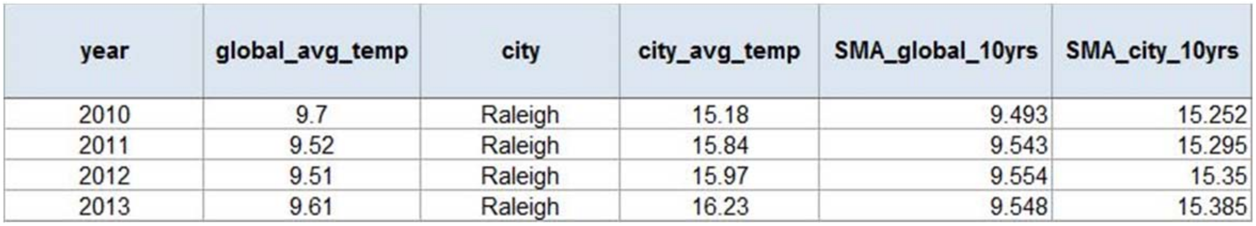
By looking at the graph above, we notice the lines are highly fluctuated and it is hard to recognize the trends. To resolve this issue, I am going to take the simple moving average for both data to smooth the lines out.

Our data are between 1750 and 2013 excluding 1780, so we have 263 years of data. To take the moving average, I will start with 10 years, i.e., having a graph for 253 entries.

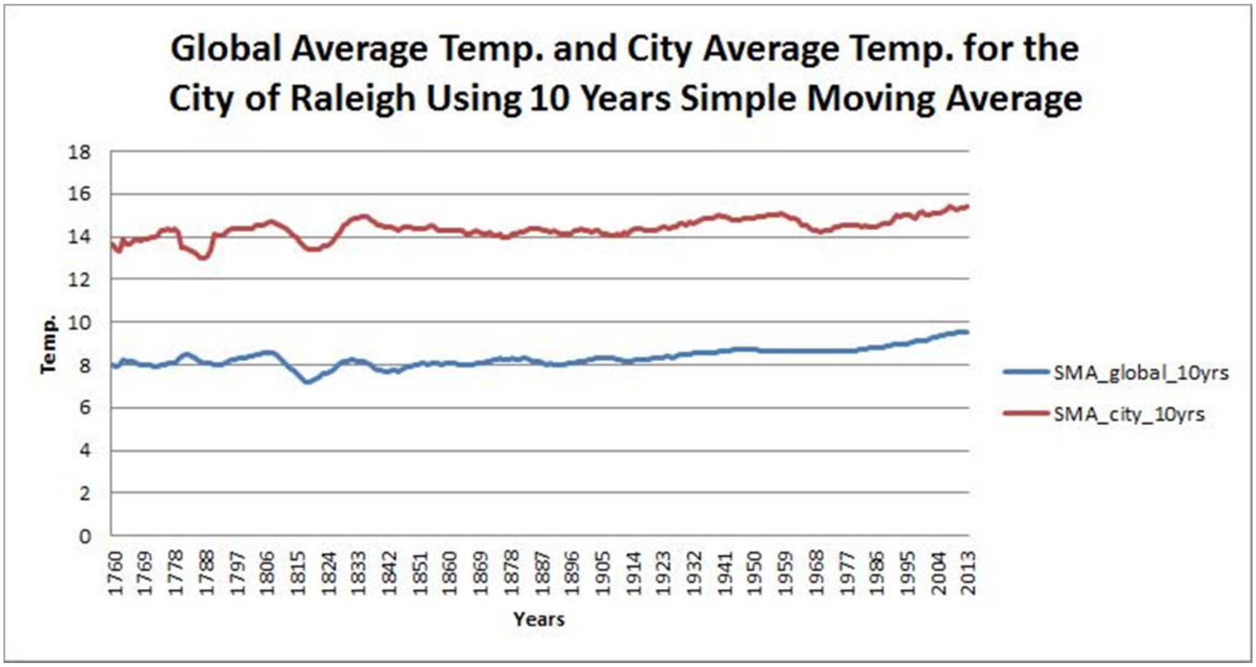


year	global_avg_temp	city	city_avg_temp	SMA_global_10yrs	SMA_city_10yrs
1750	8.72	Raleigh	15.02		
1751	7.98	Raleigh	15.79		
1752	5.78	Raleigh	8.67		
1753	8.39	Raleigh	14.41		
1754	8.47	Raleigh	14.6		
1755	8.36	Raleigh	12.02		
1756	8.85	Raleigh	14.62		
1757	9.02	Raleigh	14		
1758	6.74	Raleigh	12.96		
1759	7.99	Raleigh	13.94		
1760	7.19	Raleigh	12.6	8.03	13.603
1761	8.77	Raleigh	15.01	7.877	13.361
1762	8.61	Raleigh	14.39	7.956	13.283
1763	7.5	Raleigh	12.38	8.239	13.855
1764	8.4	Raleigh	14.38	8.15	13.652
1765	8.25	Raleigh	13.99	8.143	13.63
1766	8.41	Raleigh	14.89	8.132	13.827
1767	8.22	Raleigh	13.6	8.088	13.854
1768	6.78	Raleigh	13.31	8.008	13.814
1769	7.69	Raleigh	14.08	8.012	13.849
1770	7.69	Raleigh	13.94	7.982	13.863
1771	7.85	Raleigh	15.01	8.032	13.997
1772	8.19	Raleigh	14.6	7.94	13.997
1773	8.22	Raleigh	14.83	7.898	14.018
1774	8.77	Raleigh	14.38	7.97	14.263
1775	9.18	Raleigh	15.19	8.007	14.263
1776	8.3	Raleigh	14.09	8.1	14.383
1777	8.26	Raleigh	13.85	8.089	14.303
1778	8.54	Raleigh	11.68	8.093	14.328
1779	8.98	Raleigh	6.97	8.269	14.165
1781	8.1	Raleigh	14.57	8.398	13.454

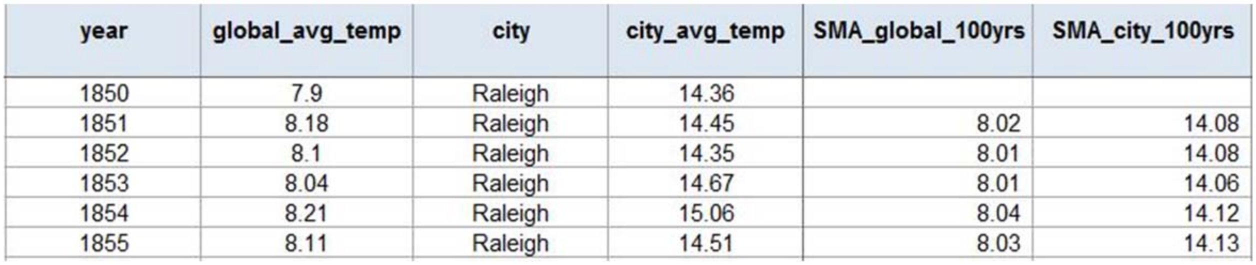
⋮



year	global_avg_temp	city	city_avg_temp	SMA_global_10yrs	SMA_city_10yrs
2010	9.7	Raleigh	15.18	9.493	15.252
2011	9.52	Raleigh	15.84	9.543	15.295
2012	9.51	Raleigh	15.97	9.554	15.35
2013	9.61	Raleigh	16.23	9.548	15.385

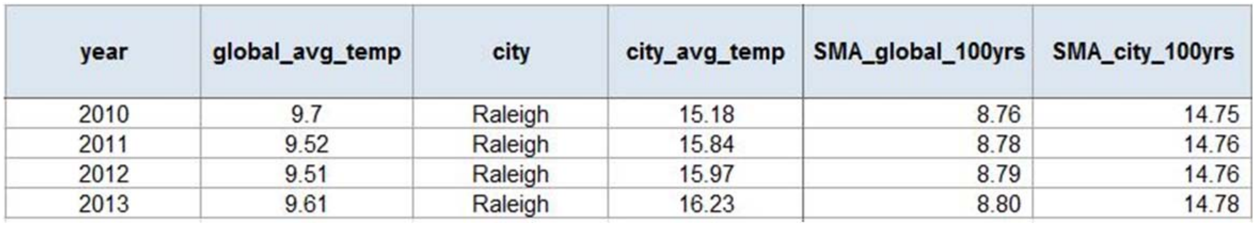


As we can see from the graph above, the lines are easier to read but there still some fluctuation. We can also notice that there are similar curves for both lines around year 1819, which was not clear in the first graph. However, although, both lines seem to have similar trend: increasing over the years, I am going to increase the number of years to 100 to smooth the line more and check the trends.

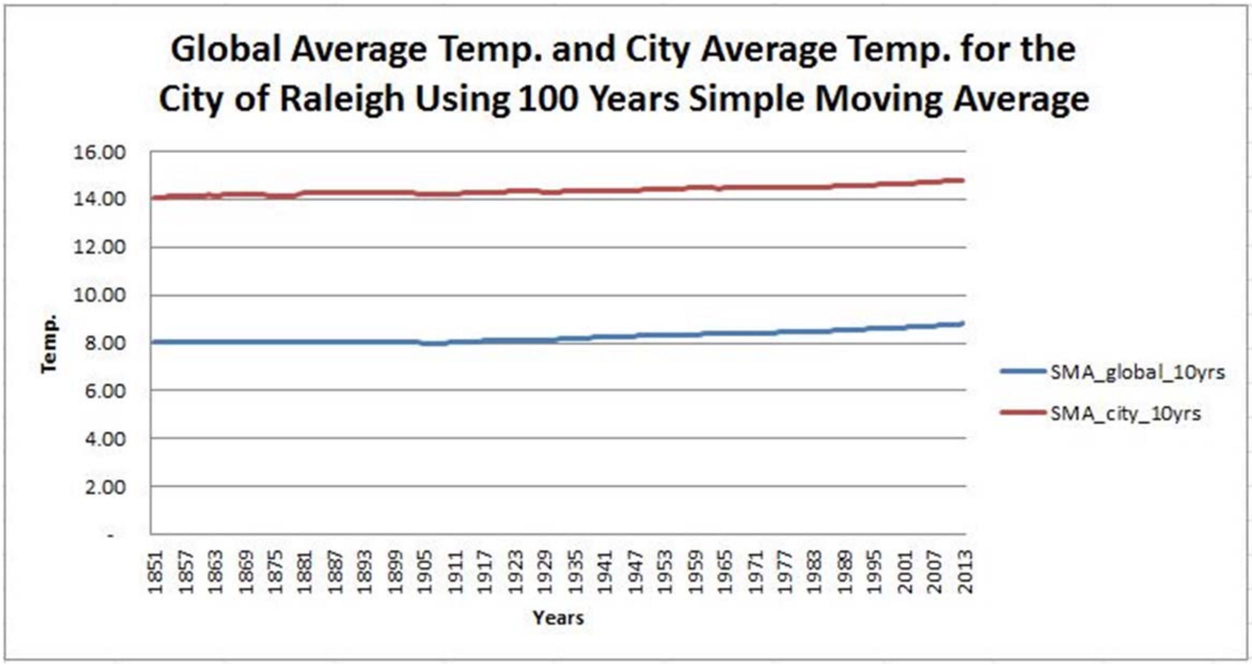


year	global_avg_temp	city	city_avg_temp	SMA_global_100yrs	SMA_city_100yrs
1850	7.9	Raleigh	14.36		
1851	8.18	Raleigh	14.45	8.02	14.08
1852	8.1	Raleigh	14.35	8.01	14.08
1853	8.04	Raleigh	14.67	8.01	14.06
1854	8.21	Raleigh	15.06	8.04	14.12
1855	8.11	Raleigh	14.51	8.03	14.13

⋮



year	global_avg_temp	city	city_avg_temp	SMA_global_100yrs	SMA_city_100yrs
2010	9.7	Raleigh	15.18	8.76	14.75
2011	9.52	Raleigh	15.84	8.78	14.76
2012	9.51	Raleigh	15.97	8.79	14.76
2013	9.61	Raleigh	16.23	8.80	14.78



From the graph above, we see the lines are smooth and the trends are clear: both lines are slightly increasing over the years.

Conclusion

After making the analysis and creating the graphs, we can conclude that the city of Raleigh is relatively hotter than the average global temperature. It also has similar trend, as both trends are slightly increasing over the years.