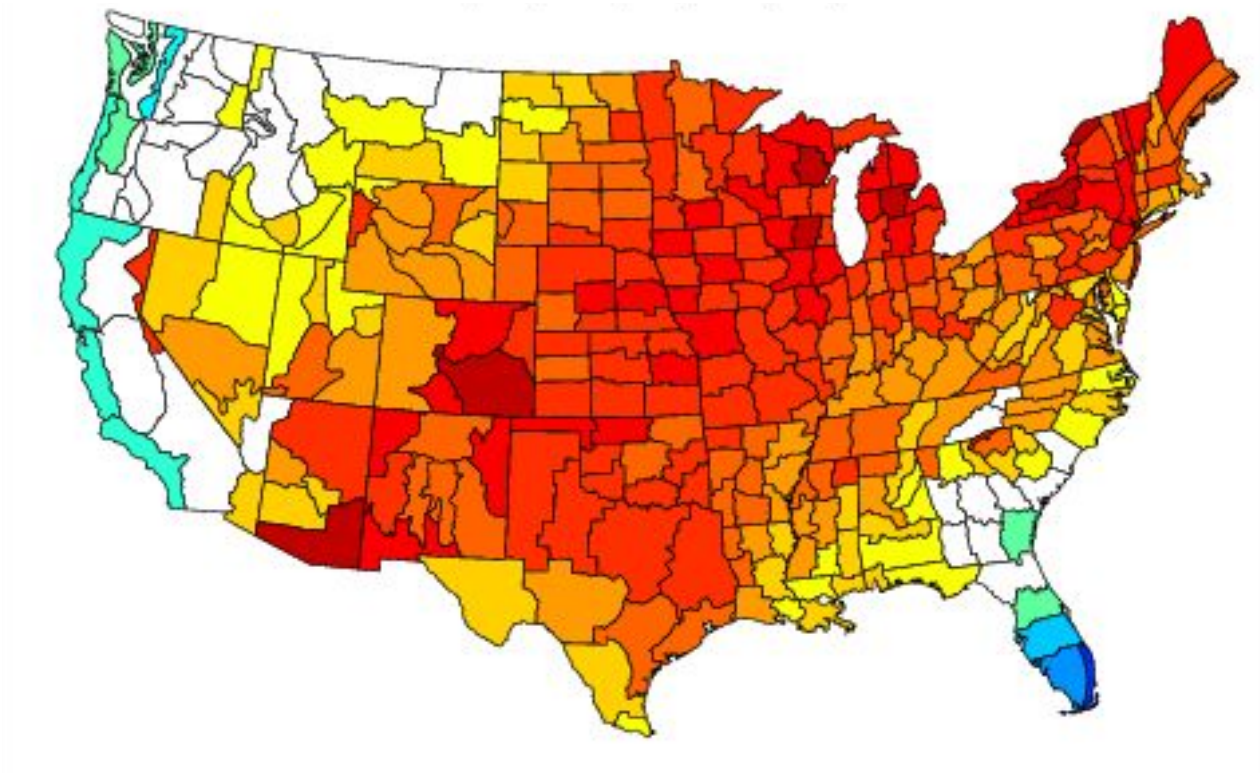
Udacity

Explore Weather Trends

Data Analyst - 1st Term

By: Nevine Gouda



Introduction

The aim of this project is to analyze local and global temperature data and compare their respective temperature trends and find any correlation or patterns between them. Where this project is an attempt to come up with visualizations and discover similarities as well as differences between the global and local (the closest city from where I live) temperature trends.

Approach

In order to achieve the project's objective; we have to do the following steps:

1. Extract the data: Where that would be done by exporting the data available in the classroom. Which can be done using simple SQL queries to extract the data and download the results as a `.csv` file(s).
2. Read the data: open up the newly created `.csv` file(s) and import them to the tool being used appropriately.
3. Calculate the moving average: By using the average temperature available globally and for each city for each year, a moving average with a window of 7 years should be calculated for both temperatures (locally and globally).
4. Plot the graph: By creating a line chart to demonstrate the trends for the local city (in my case Cairo, Egypt) and global temperatures across the years.
5. Come up with conclusions: By investigating the line charts obtained and reaching conclusions concerning trends, similarities and differences between Cairo, Egypt and globally.

Tools used

1- SQL: It is used to extract the columns and data needed from the database and load it into a `.csv` file(s).

2- Excel: It is used as an approach to load the data from the newly created `.csv` file(s), calculate the moving average, and plot the line charts.

3- Python and supporting libraries:

I. Pandas
II. Matplotlib
III. NumPy

Where they are used as another equivalent approach to excel to load the data, calculate the moving average and plot the graphs.

Implementation

In this section I will show the implementation used to solve this problem.

SQL Implementation

This section is responsible for retrieving data from the database. Where several queries were created to extract data and download them for further investigation. These queries are the following:

1- This query is responsible for finding if the city Cairo in Egypt exist in the list of cities or not. Thus this is achieved by selecting all the records from the `city_list` table that contain `Egypt` as a country **and** `Cairo` as a city.

```
1   SELECT *
2   From city_list
3   WHERE country='Egypt' AND city='Cairo';
```

Sample Output:

Output 1 results Download CSV

city	country
Cairo	Egypt

2- And since we can see that Cairo exists, now we can collect all the local city temperature data. And that can be achieved by selecting all the records from the `city_data` table that contain `Egypt` as a country **and** `Cairo` as a city.

```
1   SELECT *
2   From city_data
3   WHERE country='Egypt' AND city='Cairo';
```

Sample Output:

We can see that the city Cairo, Egypt has records in the city_data, starting from year 1808 and has 206 records of average temperature per year.

Output 206 results Download CSV

year	city	country	avg_temp
1808	Cairo	Egypt	17.11
1809	Cairo	Egypt	19.87
1810	Cairo	Egypt	19.93
1811	Cairo	Egypt	20.00
1812	Cairo	Egypt	19.93

3- Now we can investigate the global data by selecting all the records from the `global_data` table.

```
1    SELECT *
2    From global_data;
```

Sample Output:

We can observe from the output below that the table `global_data` contains 266 records where each record contains the year and its equivalent average temperature.

Output 266 results Download CSV

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

4- Therefore we can get the data from `city_data` table using the values in the `city_list` table. Where that can be done by selecting all the records from the `city_data` table where the city is **IN** another query. Where the other query selects from the `city_list` the records with `Egypt` as a country **and** `Cairo` as a city.

```
1   SELECT *
2   From city_data
3   WHERE city_data.city IN (SELECT city
4   From city_list
5   WHERE country='Egypt' AND city='Cairo');
```

Sample Output:

Output 206 results Download CSV

year	city	country	avg_temp
1808	Cairo	Egypt	17.11
1809	Cairo	Egypt	19.87
1810	Cairo	Egypt	19.93
1811	Cairo	Egypt	20.00
1812	Cairo	Egypt	19.93

5- Finally the following query is responsible for selecting the year, the global average temperature, and a city average temperature. Where the two tables `global_data` and `city_data` are joined **ON** the year. Where it only selects records with `Egypt` as a country **and** `Cairo` as a city from the `city_data` table.

```
SELECT global_data.year,
       global_data.avg_temp as global_temp,
       city_data.avg_temp as cairo_temp
From global_data JOIN city_data
ON global_data.year = city_data.year
WHERE country='Egypt' AND city='Cairo';
```

Sample Output:

Output 206 results Download CSV

year	global_temp	cairo_temp
1808	7.63	17.11
1809	7.08	19.87
1810	6.92	19.93
1811	6.86	20.00
1812	7.05	19.93

Now that we have the years that we care about as well as the local temperature, we can just download this final query result as `.csv` file to be used for the Excel and Python stages.

Please note that both Python and Excel Sheets are used to do the exact same thing. Where I chose to try both to come up with the conclusion of seeing which is easier and better for me.

Excel Implementation

Now that the data is available and downloaded as `.csv` files. We can use Microsoft Excel to calculate the moving average and start plotting. Where the moving average was calculated using excel in the following steps:

1. Open the .csv file obtained from the last SQL query with Excel
2. Create 2 new columns and name them `global_average` and `cairo_average`. As seen below in Figure 1.
3. If we want the moving average window to be 7 years, then we have to go cell `E8` and write the following equation `=AVERAGE(C2:C8)`. Where the number of values from `C2:C8 == window size == 7`.
4. Drag the equation from cell `E8` till the end of the records. And do the same for `global_average` column.

	A	B	C	D	E
1	year	global_temp	cairo_temp	global_average	cairo_average
2	1808	7.63	17.11		
3	1809	7.08	19.87		
4	1810	6.92	19.93		
5	1811	6.86	20		
6	1812	7.05	19.93		
7	1813	7.74	20.51		
8	1814	7.59	20.43	7.267142857	=AVERAGE(C2:C8)
9	1815	7.24	20.3	7.211428571	20.13857143
10	1816	6.94	20.51	7.191428571	20.23
11	1817	6.98	21.88	7.2	20.50857143
12	1818	7.83	11.6	7.338571429	19.30857143
13	1819	7.37	20.31	7.384285714	19.36285714

Figure 1

Finally a line chart can be created using columns `A`, `D` and `E`. Where this plot can be found in the results section below in Figure 3.

Python Implementation

Since the data is available and downloaded as `.csv` files. We can take advantage of that and use Python and Pandas' library as it can easily in a one liner load the csv file into a DataFrame. And using that DataFrame we can use Panda's powerful method called rolling that performs rolling window calculations based on the aggregation function provided. Thus I did the following in the attempt to implement using Python:

1. Open the .csv file obtained from the last SQL query and load it into Panda's dataFrame.
2. Call the `rolling` function on the DataFrame created with the window size set to 7 and roll on the "`year`" column. Where the aggregation function should be set to `mean`.
3. Plot the the graph on the columns updated by the rolling function.

```
# Importing the libraries needed for the project
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt

# Loading the previously created csv file using Pandas' method read_csv and load the data into Panda's DataFrames
data_frame = pd.read_csv("results.csv")
# Pand's rolling method provides rolling window calculations which is how we calculate the moving average
# Where the window is equal to the number of years we want to smooth out and calculate their average
# The window here is set to 7
rolled_df = data_frame.rolling(window=7,on="year").mean()
# Plotting the Global and Local line charts
plt.plot(data_frame['year'],rolled_df["global_temp"],label="Global")
plt.plot(data_frame['year'],rolled_df["cairo_temp"],label="Local: Cairo")
# Showing the legends for the 2 lines to be able to identify them
plt.legend()
plt.xticks(np.arange(min(data_frame['year']), max(data_frame['year'])+1, 7),rotation=90)
# Setting the plots' labels and titles
plt.xlabel("Years")
plt.ylabel("Average Temperature")
plt.title("Python: 7-Years Moving Average Temperature")
# Printing out the plots
plt.show()
# Printing some statistics
print "Min. Global Temp", min(data_frame['global_temp'])
print "Max. Global Temp", max(data_frame['global_temp'])
print "Min. Cairo Temp", min(data_frame['cairo_temp'])
print "Max. Cairo Temp", max(data_frame['cairo_temp'])
```

Results can be found in Figures 2 and 4 below.

Results & Conclusion

The statistics obtained about the maximum and minimum average temperatures can be found in Figure 2. While Figure 3 is the chart obtained from the Excel implementation mentioned before. And Figure 4 is the chart obtained from the Python implementation.

Finally, given the graphs and the experiments done on the data I have concluded the following:

1. As the size of the window for the moving average increases the smoothness of the lines relatively increases.
2. If the window size increases the range for x-axis (years) values decreases.
3. There is a huge difference between the average temperature in Cairo across the years and the global average temperature. Where Cairo, Egypt is on average higher by 10 degrees than the rest of the world.
4. The maximum global average temperature is 9.73 degrees while Cairo's maximum temperature was 23.72 degrees. With a difference of 13.99 degrees.
5. The minimum global average temperature is 6.86 degrees while Cairo's maximum temperature was 11.6 degrees. With a difference of 4.74 degrees.
6. The city of Cairo, Egypt seems to be hotter than most cities.
7. Starting the 1990's both Cairo and the global lines start to increase in temperature at a higher rate than before. Which can be a result from global warming.
8. Early in the 1800s' both cairo and global average temperature have almost similar ups and downs. However afterwards, the globe starts to have a steady and smooth curve and average temperature while Cairo still fluctuates between 20 and 22 degrees.
9. Finally, the world is getting hotter, and global warming is real and obvious. God help our mother earth.

```
Min. Global Temp 6.86
Max. Global Temp 9.73
Min. Cairo Temp 11.6
Max. Cairo Temp 23.72
```

Figure 2

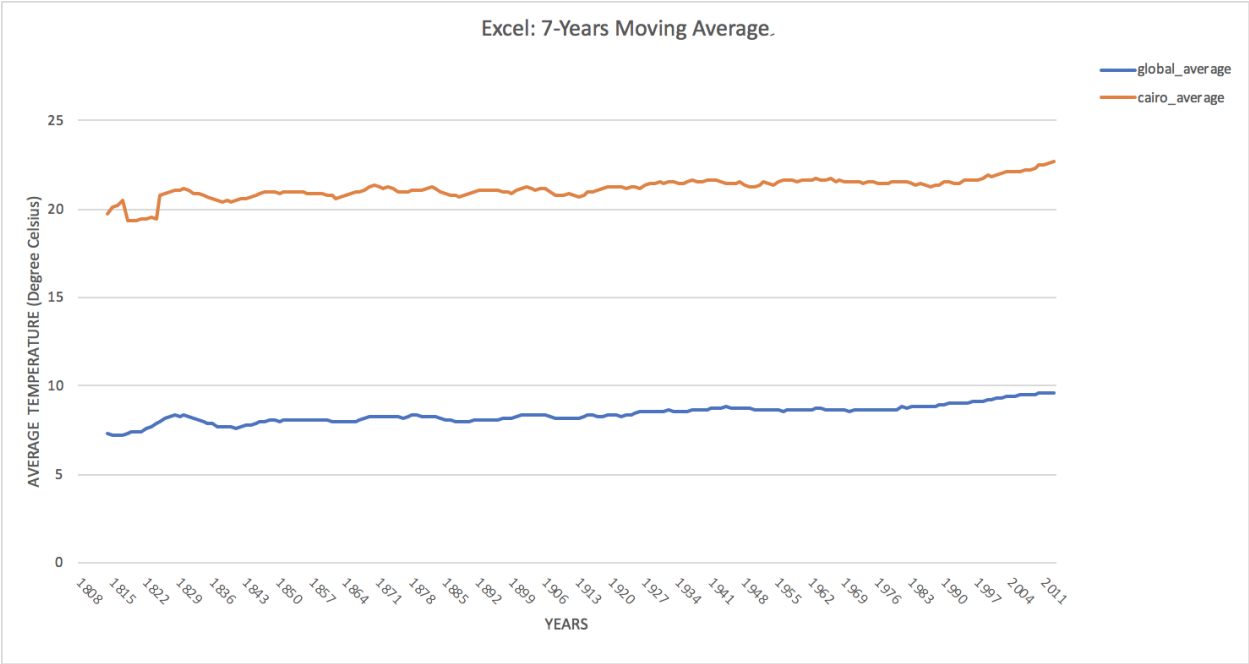


Figure 3

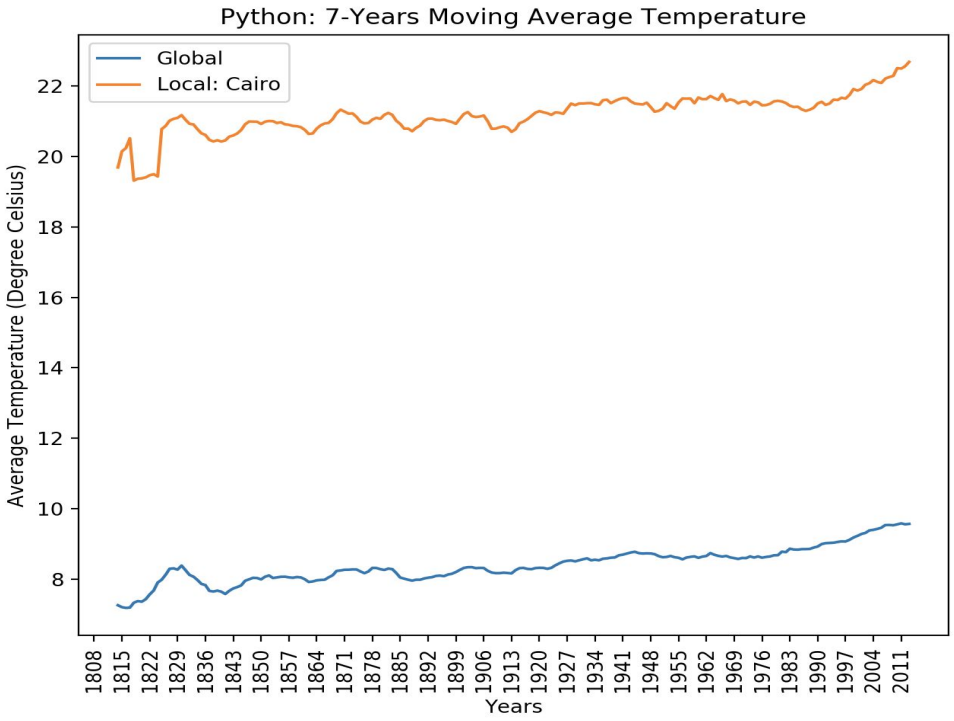


Figure 4