Explore Weather Trends

Ahmed Ali Abdelrahman
25 September 2019

- Initial Setup
    - Extract data from database
    - Install Python packages needed in the project.
    - Load both the global and city data (Cairo data)
- Read the statistics from both plot
- Initial line plot
- Plot the moving average
    - Clean the data
- Is the increase the same for city and global?

Initial Setup

Extract data from database

Download the global data as csv from the course site, using the below command:

```
SELECT *  FROM global_data
```

Do the same thing for city data, in this case, I will be using Egypt (Cairo) Data:

```
SELECT year, avg_temp  FROM city_data
WHERE country='Egypt' and city='Cairo';
```

Install Python packages needed in the project

Install and load the essential Python packages for Time Series analysis – numpy, pandas, matplotlib & seaborn.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
sns.set()
```

Load both the global and city data (Cairo data)

```
dfCairo = pd.read_csv(r"C:\Users\ARahman25\Desktop\cairo_results.csv")
dfCairo.head()
dfGlobal = pd.read_csv(r"C:\Users\ARahman25\Desktop\global_results.csv")
dfGlobal.head()
```

The last 2 commands above are used to display Cairo, Global data & shows the head of data, also no need to clean up the data since we've cleaned it using SQL during extracting from the database.

Read the statistics from both plot

dfCairo.describe()

	year	avg_temp
count	206.000000	206.000000
mean	1910.500000	21.169612
std	59.611241	0.951749
min	1808.000000	11.600000
25%	1859.250000	20.812500
50%	1910.500000	21.245000
75%	1961.750000	21.570000
max	2013.000000	23.720000

dfGlobal.describe()

	year	avg_temp
count	266.000000	266.000000
mean	1882.500000	8.369474
std	76.931788	0.584747
min	1750.000000	5.780000
25%	1816.250000	8.082500
50%	1882.500000	8.375000
75%	1948.750000	8.707500
max	2015.000000	9.830000

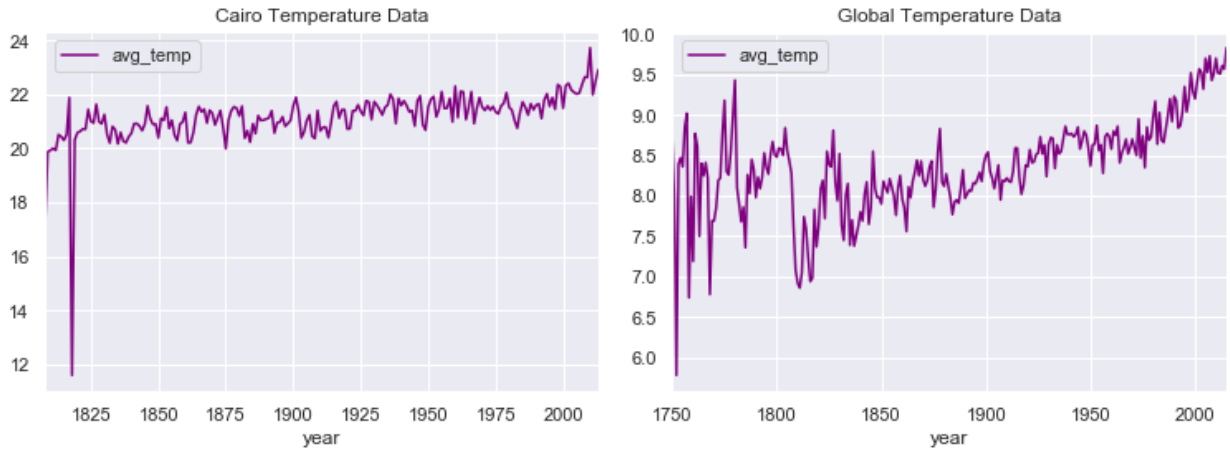
The first summary shows Cairo data while the second one shows Global data.

We can see that Cairo average temperature is always higher than global temperature, since the maximum global temperature has not reached Cairo temperature.

We can also see that global average temperature has higher range than Cairo's temperature.

Initial line plot

```
dfCairo.plot(x='year', y='avg_temp', color='purple')
dfGlobal.plot(x='year', y='avg_temp', color='purple')
plt.show()
```



From observing the above graphs, we can see that in both data there seems to be an increase of the temperature throughout the years, especially for Cairo data, also the maximum temperature of Cairo happens somewhere near 2000, yet the global temperature seems to happen at the end of the plot.

Plot the moving average

Clean the data

As we see that we've to make start & end date is the same for the both Cairo data and Global data to make measurement accurate, so here is the cleanup code for the data:

```
dfCairo2 = dfCairo[dfCairo.year >= 1850]
dfGlobal2 = dfGlobal[(dfGlobal.year >= 1850) & (dfGlobal.year <=2013)]
```

Defining the simple moving average function & applying it to our data:

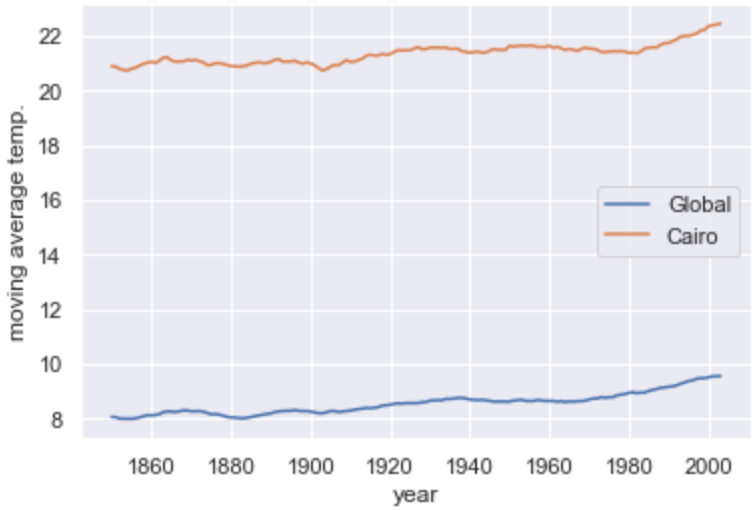
```
def SMA(values,window):
   weights = np.repeat(1.0, window)/ window
   smas = np.convolve(values, weights, 'valid')
   return smas
smaCairo = SMA(dfCairo2.avg_temp,11)
smaGlobal = SMA(dfGlobal2.avg_temp,11)
```

Rearrange the data and concatenate it to the exist data to be plotted on graphs:

```
arrCairo = np.array(smaCairo)
dfSmaCairo = pd.DataFrame(data=arrCairo.flatten())
arrGlobal = np.array(smaGlobal)
dfSmaGlobal = pd.DataFrame(data=arrGlobal.flatten())
dfCairo3 = dfCairo2.reset_index(drop=True)
dfGlobal3 = dfGlobal2.reset_index(drop=True)
dfGlobal3.drop('avg_temp', axis=1, inplace=True)
dfCairo3.drop('avg_temp', axis=1, inplace=True)
dfGlobal4 = pd.merge(dfGlobal3, dfSmaGlobal, left_index=True, right_index=True)
dfCairo4 = pd.merge(dfCairo3, dfSmaCairo, left_index=True, right_index=True)
dfCairo4.columns = ['year', 'sma']
dfGlobal4.columns = ['year', 'sma']
```

Plotting the graphs for both Cairo & Global SMAs:

```
plt.plot(dfGlobal4[('year')], dfGlobal4[('sma')], label = "line 1")
plt.plot(dfCairo4[('year')], dfCairo4[('sma')], label = "line 2")
plt.xlabel('x - axis')
plt.ylabel('y - axis')
plt.title('Two or more lines on same plot with suitable legends ')
plt.legend()
plt.show()
```



As we can see from above, both Cairo and Global temperature is increasing throughout the year.

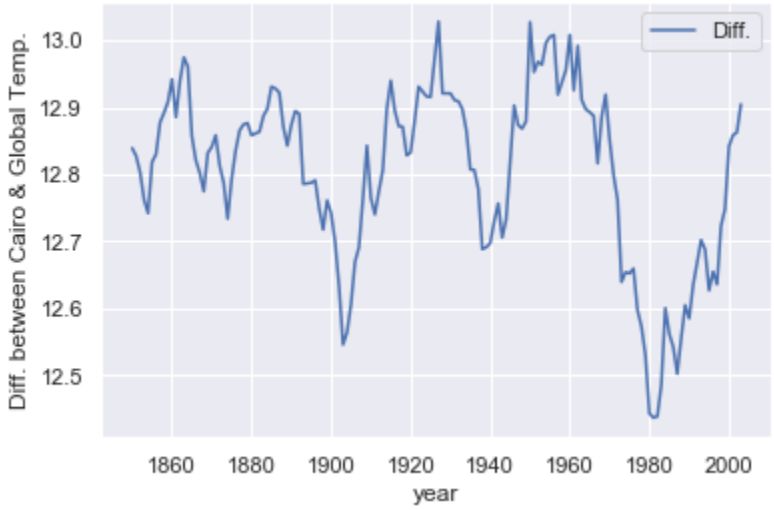
Is the increase the same for city and global?

The difference between Cairo & Global data:

```
diff = dfCairo4[('sma')] - dfGlobal4[('sma')]
```

Try to plot the data, it should be a nearly straight line if the increase is the same.

```
plt.plot(dfGlobal4[('year')], diff, label = "Diff.")
plt.xlabel('year')
plt.ylabel('Diff. between Cairo & Global Temp.')
plt.legend()
plt.show()
```



From the plot above, it is hard to tell whether the data is actually increasing or decreasing. However we can see that the difference lies around 12.8 degrees each year.

Let’s try to get the summary of the difference to find its statistics:

```
diff.describe()
```

	Difference
count	154
mean	12.80367
std	0.132141
min	12.43636
25%	12.73068
50%	12.83409
75%	12.89727
max	13.02818

From here, we can see that in the last few decades, the difference has tendency to decrease. However, it seems like both Cairo and Global temperature has increased.