First Project – Explore Weather Trends

1. SQL Data Extraction

Using the database we were provided, I was able to extract all the necessary information regarding the temperatures that were observed in both global scale and in my city Athens. In order to achiever the latter I combined both the local and the global temperatures in regard to their corresponding year, in order to create an efficient spreadsheet. The code that was used can be seen bellow:

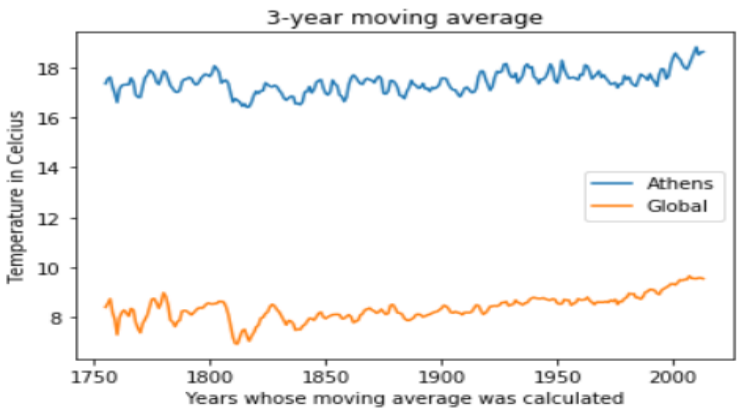
```
SELECT cd.year AS year, cd.city city, cd.avg_temp
 AVG_TEMP_ATHENS, gd.avg_temp AVG_TEMP_GLOBAL
    FROM city_data cd
    JOIN global_data gd
       ON cd.year = gd.year
    WHERE city = 'Athens'
    ORDER BY year
```

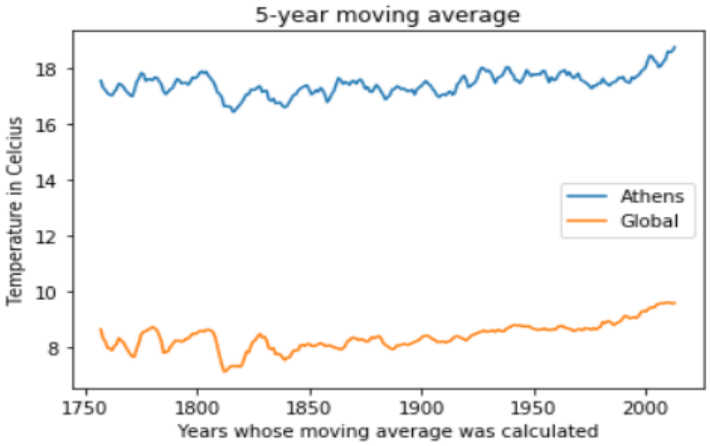
2. Moving Average Calculation

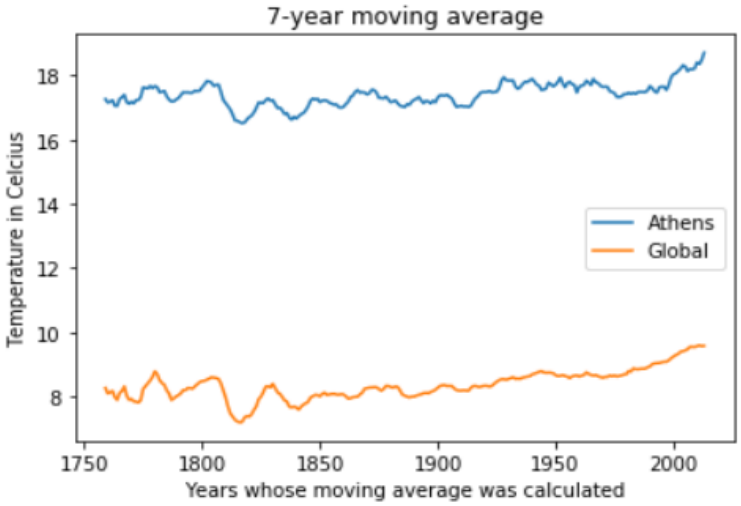
Using the pandas library, I was able to calculate the moving average for both the global temperatures and the local ones. The functions that were used were the rolling() that takes as an input the window size and extracts row values based on the window size and the mean() that calculates the mean of the passed values, in this case the ones that were extracted from the rolling() function.

```
data['global_avg_3'] = data["avg_temp_global"].rolling(window=3).mean()
data['global_avg_5'] = data["avg_temp_global"].rolling(window=5).mean()
data['global_avg_7'] = data["avg_temp_global"].rolling(window=7).mean()
data['city_avg_3'] = data["avg_temp_athens"].rolling(window=3).mean()
data['city_avg_5'] = data["avg_temp_athens"].rolling(window=5).mean()
data['city_avg_7'] = data["avg_temp_athens"].rolling(window=7).mean()
```

Bellow we can the line charts for three different window sizes (3,5,7) for both temperature measurements:


3-year moving average
Temperature in Celcius
Years whose moving average was calculated
Athens
Global
18
16
14
12
10
8
1750
1800
1850
1900
1950
2000


5-year moving average
Temperature in Celcius
Years whose moving average was calculated
Athens
Global
18
16
14
12
10
8
1750
1800
1850
1900
1950
2000

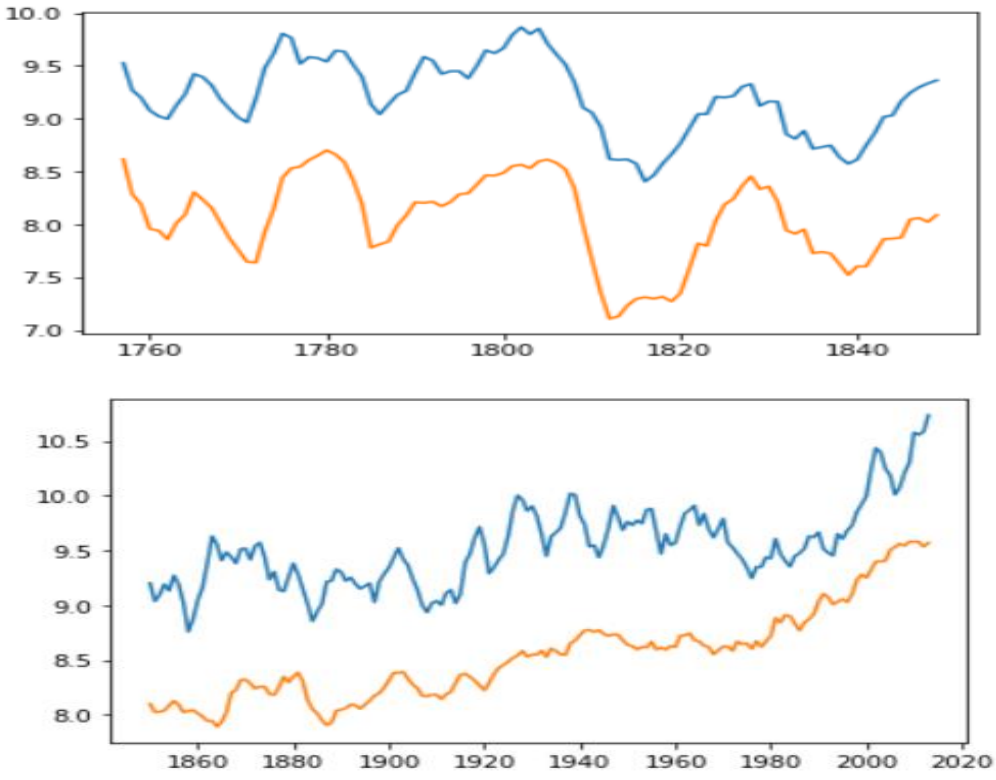

7-year moving average
Temperature in Celcius
Years whose moving average was calculated
Athens
Global
18
16
14
12
10
8
1750
1800
1850
1900
1950
2000

In my opinion I believe that the 5-year moving average depiction is the best visualization compared to the rest. Firstly, the important patterns have not been filtered and are easy to observe. Secondly, it is easy to spot the general trends since a lot of noise has been removed. On the contrary, the 3-year moving average contains a lot of noise and is hard to distinguish important patterns and general trends. Lastly, the 7-year moving average lacks a lot of important patterns, since they have been filtered.

3. Observations

- The diagram can be split in two periods, one until the year 1850 and one after that year.

- During the first period, both the global temperature and the temperature in Athens were subject to minor temperature fluctuations. In the second period it can been seen that there was a slight gradual increase in both observed temperatures. Furthermore, it can be observed that both temperature line graphs have similar trends and patterns. These two periods can be seen bellow:



- Another small detail that should be noted is that in the cases were a peak or a minimum was observed in global temperature, a similar one was observed in the local temperature.

- Afterwards, the averages of both temperatures were calculated for every 50 years, in order to make comparisons between the global temperature and the one in Athens.

In [16]:
```
avg_global = []
avg_athens = []
years = []
difference = []
i = 0
year = 1753


for j in range(5): #average per 50 years calculation


    if(i<200):
        avg_global.append(data.iloc[i:(i+49),5].dropna().mean())
        avg_athens.append(data.iloc[i:(i+49),8].dropna().mean())
        years.append(str(year)+"-"+str(year+49))
        difference.append(avg_athens[j]-avg_global[j])

    else:
        avg_global.append(data.iloc[i:,5].dropna().mean())
        avg_athens.append(data.iloc[i:,8].dropna().mean())
        years.append(str(year)+"-"+str(2013))
        difference.append(avg_athens[j]-avg_global[j])

    i = i + 50
    year = year + 50
```

In [17]:
```
df = pd.DataFrame(list(zip(years, avg_athens, avg_global,difference)), columns=["Years", "Athens", "Gl
obal", "Difference"])
```

	Years	Athens	Global	Difference
0	1753-1802	17.387333	8.210089	9.177244
1	1803-1852	17.016286	7.876735	9.139551
2	1853-1902	17.248816	8.132980	9.115837
3	1903-1952	17.531388	8.477388	9.054000
4	1953-2013	17.767934	8.941869	8.826066

- It can be clearly seen that Athens is hotter compared to the global average.

- The difference between the average temperature in Athens and the global average temperature has remained roughly the same at 8-9 degrees.

- It can be clearly seen from the line graph that for the years between 1750 and 1850 the temperature was fluctuating, until it reached a minimum at approximately 1830. Afterwards, the global temperature, as well as the temperature in Athens, have been gradually increasing. Seeing that pattern, it can be safely assumed that the world is indeed getting hotter.

- A rough estimation of that increase would be that the temperature is increasing by 0.5 every 50 years or by 1 every 100 year.

- It should also be noted that until 1850, increased temperatures used to be compensated by low temperatures, thus keeping the average steady in the long term. It is safe to assume that temperatures used to cycle between low and hot. That pattern seems to have been disturbed, after the year 1850 temperature has been increasing without decreasing after a given period of time, as it used to before.

- Previously, it was stated that the temperature in Athens had a similar trend with the global temperature. That can be confirmed from the correlation coefficient, which shows a strong positive almost linear relationship.

	global_avg_5	city_avg_5
global_avg_5	1.000000	0.866516
city_avg_5	0.866516	1.000000

- Using the data that I was provided, I was able to create a linear regression model that can predict the temperatures in Athens, given the global temperature.

The python notebook that was used for this project can be found at the following https://github.com/venetisgr/Udacity-Data-Analyst-Nanodegree