Udacity Data Analyst Nanodegree

Project 01

Explore weather trends

Image source: https://www.visitraleigh.com/meetings-and-conventions/blog/post/destination-news-fall-2018/

Name: Hao Cui

Introduction:

I compared the temperature trends of Raleigh, North Carolina with the global temperature trends. I extracted and downloaded the csv data by using SQL query and analyzed the data by using Jupyter notebook with python 3.6. Overall, I find both global and local temperature has been increasing from the time period I analyzed (1750-2013) and there is an obvious increase trends especially after 1850(the start of the 2nd industry revolution).

Outline:

Step1:

I extracted and downloaded the data of both global and local temperature by using SQL query statements(Figure1, figure2)

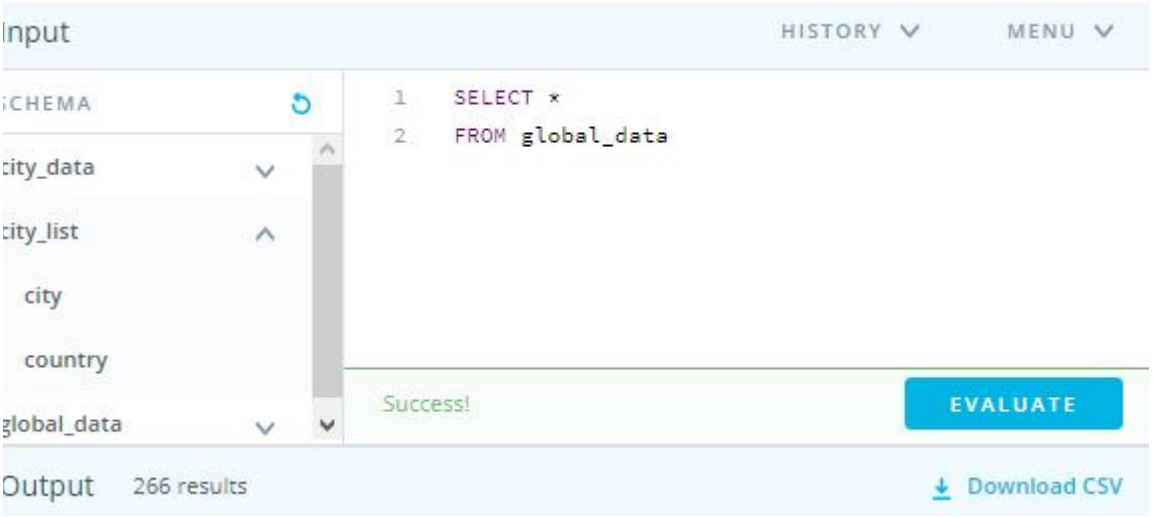

Figure1: extract and download global temperature data

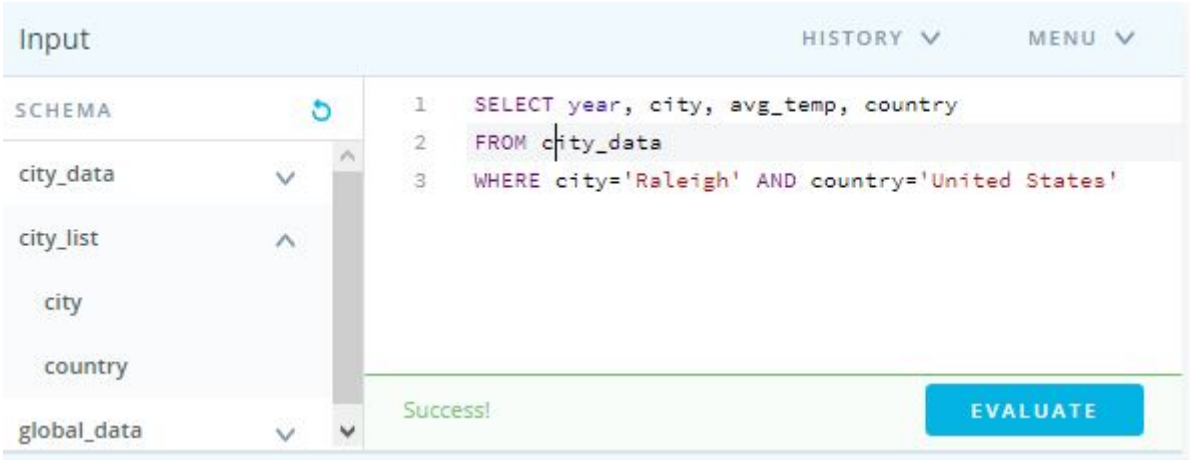

Figure2: extract and download local temperature data

Step2: I use the Jupyter Notebook to read and import the downloaded csv data.(figure3)

```
In [52]:
1 data1=pd.read_csv('global_data.csv')
2 data2=pd.read_csv('raleigh_data.csv')
```

Figure3: code for reading and importing the downloaded csv data

Step3: After check the 1st 5 rows and last 5 rows of the data to understand what columns are in the data frame.

```
1 print(data1.head()), print(data1.tail())
```

```
   year  avg_temp
0  1750      8.72
1  1751      7.98
2  1752      5.78
3  1753      8.39
4  1754      8.47
     year  avg_temp
261  2011      9.52
262  2012      9.51
263  2013      9.61
264  2014      9.57
265  2015      9.83

(None, None)
```

```
1 print(data2.head(), data2.tail())
```

```
   year     city  avg_temp        country
0  1743  Raleigh      7.81  United States
1  1744  Raleigh     16.02  United States
2  1745  Raleigh      7.61  United States
3  1746  Raleigh       NaN  United States
4  1747  Raleigh       NaN  United States      year     city  avg_temp
266  2009  Raleigh     14.90  United States
267  2010  Raleigh     15.18  United States
268  2011  Raleigh     15.84  United States
269  2012  Raleigh     15.97  United States
270  2013  Raleigh     16.23  United States
```

Step4: Based on two dataframes available, I decided to cleaning and processing the data for further analysis

```
data3=data2.drop(columns=['city','country'])
data3.columns=['year','avg_temp_r']
data3.head()
```

Figure4: code for drop the columns and change the name of the columns.

Step5: Using pandas.merge to merge the two dataframe into one dataframe.

In order to use the rolling method, I forward the na data in the dataframe based on the history records from websites. After that, I used the rolling method to calculate the 10 days moving average for both global and local average temperature.

```
#Merge dataframe 1 and 3
df_combine=pd.merge(data1, data3)
# in order to get a complete timeseries
# forward-fill the na data in the dataframe(based on history records)
df_combine.fillna(method='ffill', inplace=True)
#Rolling 10 MA
rolling=df_combine[['avg_temp_g','avg_temp_r']].rolling(10, center=True).mean()
rolling['year']=(df_combine['year']+5)
df_combine.iloc[30]
```

Figurer5: step5

Step6: By using objective oriented method, I drew the average trends and rolling trends with tools from matplotlib libraries.

```
fig,ax=plt.subplots(2, facecolor='lightgray', figsize=(13,8), sharex=True)
ax[0].plot(df_combine['year'], df_combine['avg_temp_g'],label=['global'])
ax[0].plot(df_combine['year'], df_combine['avg_temp_r'],label=['raleigh'])
ax[1].plot(rolling['year'], rolling['avg_temp_g'],label=['global'])
ax[1].plot(rolling['year'], rolling['avg_temp_r'],label=['raleigh'])
ax[0].set(title='Annual average global vs.Raleigh temperature (1750-2013)',
          ylabel='avg. temperature(Degree Celsius)')
ax[1].set(title='Rolling 10 MA global vs.Raleigh temperature (1750-2013)',
          ylabel='avg. temperature(Degree Celsius)')
ax[0].legend(loc='lower right')
ax[1].legend(loc='lower right')
ax[0].annotate('Hardest Winters', xy=(1780,7), xycoords='data',
               bbox=dict(boxstyle='round',fc='none',ec='gray'), xytext=(20,-20),
               textcoords='offset points', ha='center', arrowprops=dict(arrowstyle='->'))
ax[0].axvline(1850, color='r', linestyle='--')
ax[1].axvline(1850, color='r', linestyle='--')
```

Figure6: step 6

Consideration: I try to make a comparison between the global average temperature and local temperature and another comparison with rolling 10 days average temperature. Thus, I combined all the visualization graphs into one graph which facilitate such comparison.

Line chart for my visualization

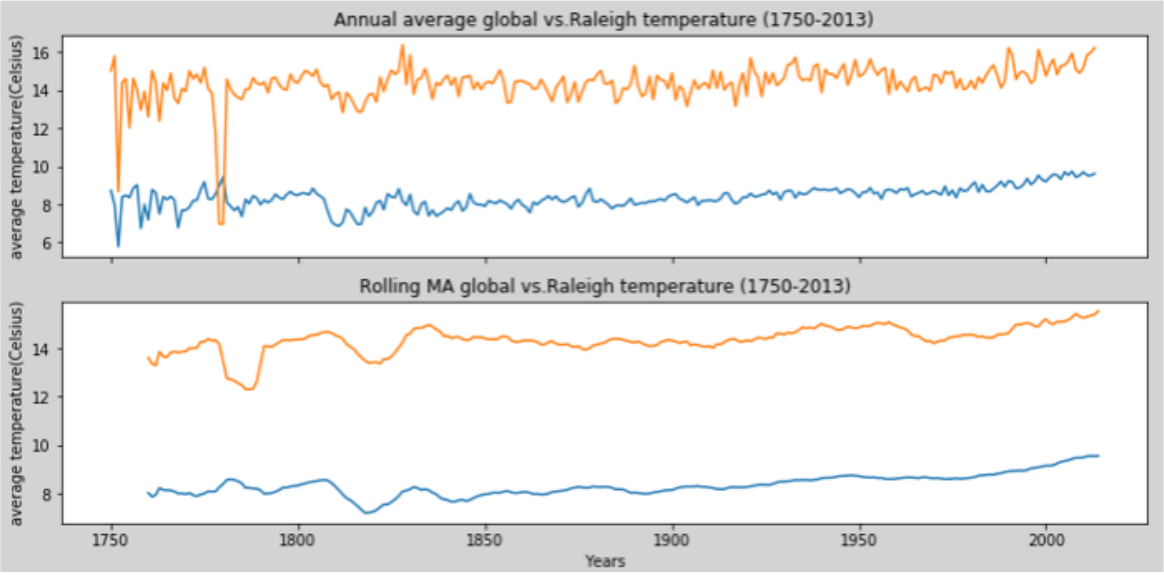

Figure7: Line chart for my visualization

Observation:

- Overall, the global average temperatures are higher than the local average temperature (Raleigh) where the global temperature ranges from 6-8 degrees Celsius and the Raleigh temperature ranges from 9-16(excluding extreme case)
- There is an outlier in the Raleigh temperature trend which is in 1779 and was described as one of the hardest winters recorded by history.
- Based on both rolling 10 MA graphs and trends graphs, there is a short dip in temperature from 1800 to 1830
- Both graphs show that the temperatures have been increasing since 1850(the 2nd industry revolution).
- Although overall trends of the Raleigh is increasing, there is a short decrease from 1950 to 1970 based on the rolling moving average graph.

Conclusion:

Based on the available data, we can see an increase trends in terms of global temperature.

Code reference:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
data1=pd.read_csv('global_data.csv')
data2=pd.read_csv('raleigh_data.csv')
print(data1.head()), print(data1.tail())
data1.columns=['year','avg_temp_g']
data3=data2.drop(columns=['city','country'])
data3.columns=['year','avg_temp_r']
#Merge dataframe 1 and 3
df_combine=pd.merge(data1, data3)
# in order to get a complete timeseries
# forward-fill the na data in the dataframe(based on history records)
df_combine.fillna(method='ffill', inplace=True)
#Rolling 10 MA
rolling=df_combine[['avg_temp_g','avg_temp_r']].rolling(10, center=True).mean()
rolling['year']=(df_combine['year']+5)
df_combine.iloc[30]
fig,ax=plt.subplots(2, facecolor='lightgray', figsize=(13,8), sharex=True)
ax[0].plot(df_combine['year'], df_combine['avg_temp_g'],label=['global'])
ax[0].plot(df_combine['year'], df_combine['avg_temp_r'],label=['raleigh'])
ax[1].plot(rolling['year'], rolling['avg_temp_g'],label=['global'])
ax[1].plot(rolling['year'], rolling['avg_temp_r'],label=['raleigh'])
ax[0].set(title='Annual average global vs.Raleigh temperature (1750-2013)',
      ylabel='avg. temperature(Degree Celsius)')
ax[1].set(title='Rolling 10 MA global vs.Raleigh temperature (1750-2013)',
      ylabel='avg. temperature(Degree Celsius)')
ax[0].legend(loc='lower right')
ax[1].legend(loc='lower right')
ax[0].annotate('Hardest Winters', xy=(1780,7), xycoords='data',
          bbox=dict(boxstyle='round',fc='none',ec='gray'), xytext=(20,-20),
          textcoords='offset points', ha='center', arrowprops=dict(arrowstyle='->'))
ax[0].axvline(1850, color='r', linestyle='--')
ax[1].axvline(1850, color='r', linestyle='--')
```