Weather_Trends

July 31, 2019

1 DAND Project 1: Explore Weather Trends

In this project, I will analyze local and global temperature data and compare the temperature trend of my city, Hong Kong, with the global counterpart.

1.1 Table of Contents

Introduction

During the data gathering process, I used SQL queries to extract the temperature data for both the world and Hong Kong from the Udacity's database, and export them into a CSV file format. Then, I will use Pandas, instead of a spreadsheet, to load the CSV files becasue of its ease in assessing and cleaning data.

After data wrangling, I will use Matplotlib to create line charts to visualize the temperatures in both the globe and Hong Kong for the sake of comparison. Before plotting the moving averages, I will first plot the yearly averages in order to show the smoothing effects, by comparing the plots on yearly averages with the plots on moving averages. Moreover, I will calculate the moving averages in both 5- and 10-year windows - dropping the data from the first four or nine years and take the average of the temperatures in the current and its previous four or nine years.

Given the moving averages calculated from two time period windows, I can choose which one could make the trends more observable in the visualization. In addition to making trends more observable, I would also like to do a direct comparison between the local and global temperature trends in the visualization. Based on the data visualization, I will at the end draw some observations about the overall temperature trends in the world and Hong Kong, as well as their similarities and differences.

Data Wrangling

After gathering the temperature data for both the world and Hong Kong by writing SQL queries in the Udacity's database, I export the two pieces of data into a CSV file format. Pandas is then used for loading the CSV files for the purposes of accessing and cleaning the data.

[2]:
```
# Import pandas library
import pandas as pd
```

[3]:
```python
# Load the Hong Kong temperature data
hk_temp = pd.read_csv("hk_temp.csv")
hk_temp.head()
```

[3]:

	year	avg_temp
0	1840	23.71
1	1841	20.76
2	1842	20.96
3	1843	21.05
4	1844	20.66

[4]:
```python
# Load the global temperature data
global_temp = pd.read_csv("global_temp.csv")
global_temp.head()
```

[4]:

	year	avg_temp
0	1750	8.72
1	1751	7.98
2	1752	5.78
3	1753	8.39
4	1754	8.47

[5]:
```python
# Print the shapes of the two dataframes
print("The shape of the HK temperature dataframe:", hk_temp.shape)
print("The shape of the global temperature dataframe:", global_temp.shape)
```

```
The shape of the HK temperature dataframe: (174, 2)
The shape of the global temperature dataframe: (266, 2)
```

The main difference between the two data sets is the number of entries: there are 174 rows for the Hong Kong temperature data set and 266 rows for the global one. If we take a look at the data sets themselves, the different numbers of entries in the data sets are due to the different initial years of record in the city and the globe.

[6]:
```python
# Check the types of the columns of hk_temp dataframe
hk_temp.dtypes
```

[6]:
```
year          int64
avg_temp    float64
dtype: object
```

[7]:
```python
# Check the types of the columns of global_temp dataframe
global_temp.dtypes
```

[7]:
```
year          int64
avg_temp    float64
dtype: object
```

[8]:
```python
# Check whether there is any null value in the columns of hk_temp dataframe
hk_temp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 174 entries, 0 to 173
```

```
Data columns (total 2 columns):
year        174 non-null int64
avg_temp    174 non-null float64
dtypes: float64(1), int64(1)
memory usage: 2.8 KB
```

[9]:
```python
# Check whether there is any null value in the columns of global_temp dataframe
global_temp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 266 entries, 0 to 265
Data columns (total 2 columns):
year        266 non-null int64
avg_temp    266 non-null float64
dtypes: float64(1), int64(1)
memory usage: 4.2 KB
```

[10]:
```python
# Check whether there is any duplicated value in the columns of hk_temp␣
 ↪dataframe
sum(hk_temp.duplicated())
```

[10]: 0

[11]:
```python
# Check whether there is any duplicated value in the columns of global_temp␣
 ↪dataframe
sum(global_temp.duplicated())
```

[11]: 0

As the above results show, the data types for the year and temperature are correct: integer for the year and float for the temperature. Moreover, there is neither missing values nor duplicates in the data sets. According to Hadley Wickham's definition, (1) each variable forms a column, (2) each observation forms a row, and (3) each type of observational unit forms a table, in a tidy data set. These two data sets are thus tidy, and they also do not have any quality issues; hence, there is no effort needed for cleaning the data.

To assess and build an intuition on the data, basic descriptive statistics are obtained by using the describe() function.

[12]:
```python
# Obtain the five-number summary on the hk_temp dataframe
hk_temp.describe()
```

[12]:

	year	avg_temp
count	174.000000	174.000000
mean	1926.500000	21.430862
std	50.373604	0.512762
min	1840.000000	20.170000
25%	1883.250000	21.072500
50%	1926.500000	21.430000
75%	1969.750000	21.782500
max	2013.000000	23.710000

[13]:
```python
# Obtain the five-number summary on the global_temp dataframe
global_temp.describe()
```

[13]:
```
             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
```

As we can see from above, the two measures of center, mean and median, are similar to each other in both of the data sets, but the two measures between the two data sets are very different: 21.43° in Hong Kong and 8.37° in the globe. The measures of spread in both data sets are shown as follows:

[14]:
```python
print("The range of the temperature data in Hong Kong: ",
      hk_temp.describe()["avg_temp"]["max"] - hk_temp.
 ↪describe()["avg_temp"]["min"])
print("The range of the temperature data in the world: ",
      global_temp.describe()["avg_temp"]["max"] - global_temp.
 ↪describe()["avg_temp"]["min"])
print("The interquartile range of the temperature data in Hong Kong: ",
      hk_temp.describe()["avg_temp"]["75%"] - hk_temp.
 ↪describe()["avg_temp"]["25%"])
print("The interquartile range of the temperature data in the world: ",
      global_temp.describe()["avg_temp"]["75%"] - global_temp.
 ↪describe()["avg_temp"]["25%"])
print("The standard deviation of the temperature data in Hong Kong: ",
      hk_temp.describe()["avg_temp"]["std"])
print("The standard deviation of the temperature data in the world: ",
      global_temp.describe()["avg_temp"]["std"])
print("The variance of the temperature data in Hong Kong: ",
      hk_temp.describe()["avg_temp"]["std"]**2)
print("The variance of the temperature data in the world: ",
      global_temp.describe()["avg_temp"]["std"]**2)
```

```
The range of the temperature data in Hong Kong:  3.539999999999999
The range of the temperature data in the world:  4.05
The interquartile range of the temperature data in Hong Kong:
0.7100000000000044
The interquartile range of the temperature data in the world:  0.625
The standard deviation of the temperature data in Hong Kong:  0.5127618219674421
The standard deviation of the temperature data in the world:  0.5847474097994193
The variance of the temperature data in Hong Kong:  0.2629246860673708
The variance of the temperature data in the world:  0.34192953326713
```

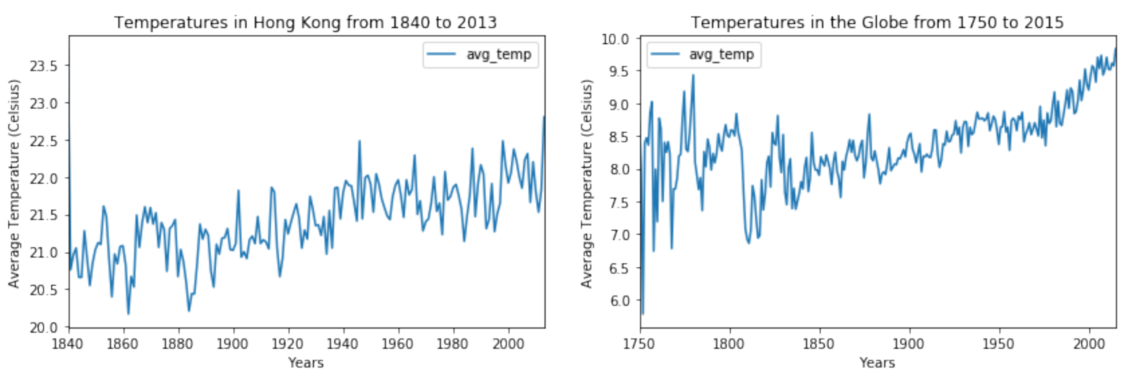
Data Visualization

Before plotting the moving average temperatures of both data sets, I would like to first plot a line chart of the yearly average temperatures as a baseline for comparison to see the smoothing effects. I will then calculate the two moving averages by taking the average of temperatures of (1) the current year and its previous four years as well as (2) the current year and its previous nine years, and plot line charts on them.

[15]:
```
# Import matplotlib library
import matplotlib.pylab as plt
%matplotlib inline
```

Part I: Line Chart of Yearly Average Terperatures (Baseline Chart for Comparison)

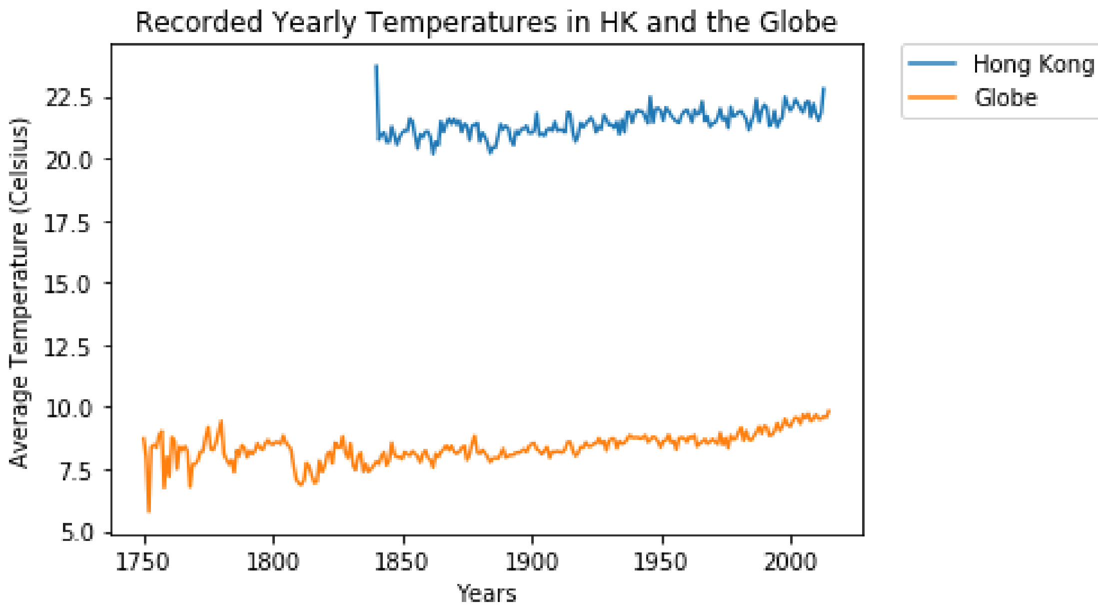
[16]:
```
# Plot the two sets of data along with each other
fig, axes = plt.subplots(1, 2, figsize=(14, 4), squeeze=False)
hk_temp.plot(x = "year", y = "avg_temp", ax = axes[0, 0], kind = "line",
             title = "Temperatures in Hong Kong from 1840 to 2013")
global_temp.plot(x = "year", y = "avg_temp", ax = axes[0, 1], kind = "line",
                 title = "Temperatures in the Globe from 1750 to 2015")
axes[0, 0].set_xlabel("Years")
axes[0, 0].set_ylabel("Average Temperature (Celsius)")
axes[0, 1].set_xlabel("Years")
axes[0, 1].set_ylabel("Average Temperature (Celsius)")
plt.show()
```



These two plots are line charts on the yearly average temperatures for the world and Hong Kong. Since both charts have different x and y axes, it would be hard to do a rough comparison between the globe and Hong Kong, and thus I plot them on the same scale with different colors in the following:

[17]:
```
# Plot the two sets of data on the same scale
plt.subplot(111)
plt.plot(hk_temp["year"], hk_temp["avg_temp"], label="Hong Kong")
plt.plot(global_temp["year"], global_temp["avg_temp"], label="Globe")
plt.title("Recorded Yearly Temperatures in HK and the Globe")
plt.xlabel('Years')
plt.ylabel('Average Temperature (Celsius)')
```

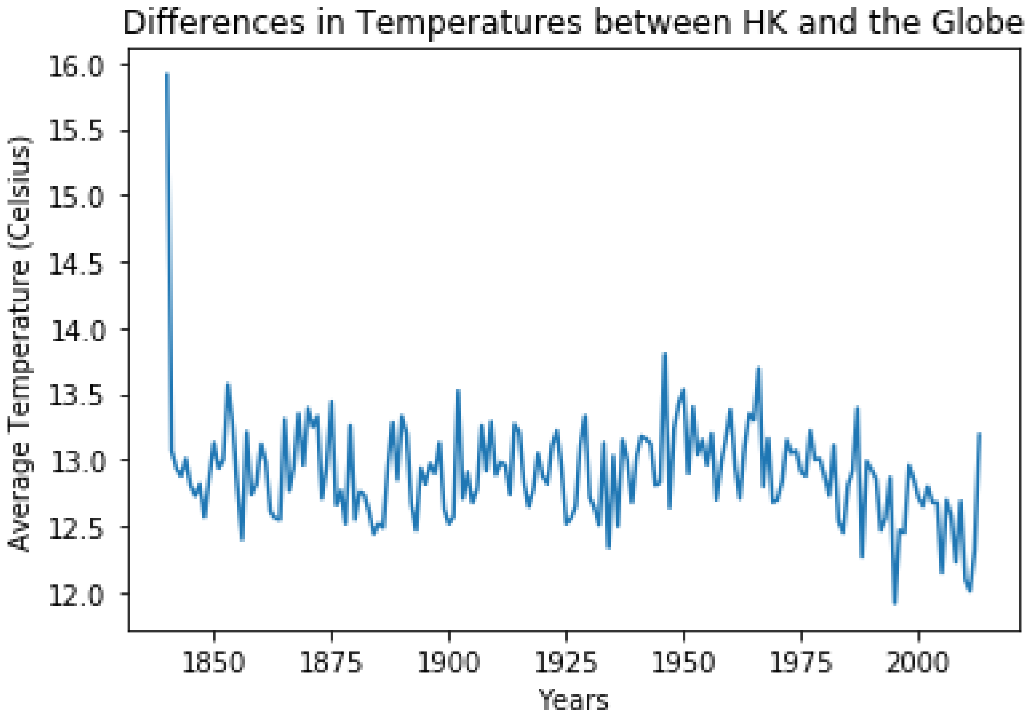
```
# Place a legend to the right of this smaller subplot
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
plt.show()
```



The above plot shows two things: (1) Hong Kong temperature data was recorded later than the global one for about a hundred years; (2) Hong Kong temperatures are higher than the global temperatures, which makes sense because the city is in a humid subtropical climate zone characterized by hot and humid summers as well as cool to mild winters. However, it is still very hard to visually compare the trends of Hong Kong and the world, given the fluctuations of the temperature data and their different scales for Hong Kong and the globe.

To have an idea about the changes in temperatures over time in Hong Kong as compared with those changes in the world, a line chart on the differences in temperatures between the two can be drawn in the period during which both the temperatures in Hong Kong and the globe were recorded.

[18]:
```
# Obtain the global data subset with the same years of records as the HK data␣
↪set
global_temp_subset = global_temp[(global_temp["year"] >= min(hk_temp["year"])) &
                                 (global_temp["year"] <= max(hk_temp["year"]))]
global_temp_subset.reset_index(inplace = True)
# Plot the difference in temperatures between HK and the globe
plt.plot(hk_temp["year"],
         hk_temp["avg_temp"] - global_temp_subset["avg_temp"])
plt.xlabel('Years')
plt.ylabel('Average Temperature (Celsius)')
plt.title("Differences in Temperatures between HK and the Globe")
plt.show()
```



As we can see from the above chart, the difference in temperatures between Hong Kong and the world did not rise during the period. Yet, line charts should be ploted on the moving average temperatures to make the trend more observable.

Part II - Line Chart of Moving Average Temperatures (5-Year Window) In this part, I will plot the same charts as in part I, except that I will plot the moving average temperatures in a five-year window instead of the yearly average temperatures, so that we can see the smoothing effects where the trends in temperature become clearer.

[19]:
```
# Obtain the moving average temperatures in 5-year window for hk_temp dataframe
hk_temp_rolling_5 = hk_temp["avg_temp"].rolling(5).mean()
hk_mov_avg_year_5 = hk_temp["year"][hk_temp_rolling_5.dropna().index].
 ↪reset_index(drop=True)
hk_mov_avg_temp_5 = pd.Series(hk_temp_rolling_5.dropna().values).
 ↪rename("mov_avg")

#hk_mov_avg_year_5 = pd.Series([])
#hk_mov_avg_temp_5 = pd.Series([])
#for i in range(len(hk_temp["avg_temp"]) - 4):
#    hk_mov_avg_year_5[i] = hk_temp["year"][i+4] # Show the fifth year of␣
 ↪moving averages
#    hk_mov_avg_temp_5[i] = (hk_temp["avg_temp"][i:i+5]).mean() # Calculate the␣
 ↪moving averages
```

[20]:
```
# Combine the two Series of years and moving average temperatures into a␣
↪dataframe
hk_mov_avg_5 = pd.concat([hk_mov_avg_year_5, hk_mov_avg_temp_5], axis = 1)
hk_mov_avg_5.rename(columns = {0: "year", 1: "mov_avg"}, inplace = True) #␣
↪Rename the columns
hk_mov_avg_5.head() # Use .tail() to see if it works
```

[20]:

	year	mov_avg
0	1844	21.428
1	1845	20.818
2	1846	20.922
3	1847	20.912
4	1848	20.812

[21]:
```
# Obtain the moving average temperatures in 5-year window for global_temp␣
↪dataframe
global_temp_rolling_5 = global_temp["avg_temp"].rolling(5).mean()
global_mov_avg_year_5 = global_temp["year"][global_temp_rolling_5.dropna().
↪index].reset_index(drop=True)
global_mov_avg_temp_5 = pd.Series(global_temp_rolling_5.dropna().values).
↪rename("mov_avg")

#global_mov_avg_year_5 = pd.Series([])
#global_mov_avg_temp_5 = pd.Series([])
#for i in range(len(global_temp["avg_temp"]) - 4):
#    global_mov_avg_year_5[i] = global_temp["year"][i+4] # Show the fifth year␣
↪of moving averages
#    global_mov_avg_temp_5[i] = (global_temp["avg_temp"][i:i+5]).mean() #␣
↪Calculate the moving averages
```

[22]:
```
# Combine the two Series of years and moving average temperatures into a␣
↪dataframe
global_mov_avg_5 = pd.concat([global_mov_avg_year_5, global_mov_avg_temp_5],␣
↪axis = 1)
global_mov_avg_5.rename(columns = {0: "year", 1: "mov_avg"}, inplace = True) #␣
↪Rename the columns
global_mov_avg_5.tail() # Use .tail() to see if it works
```

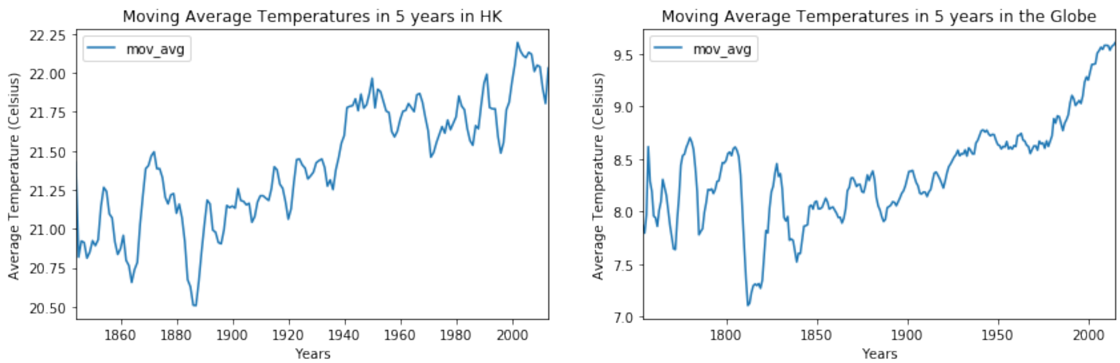
[22]:

	year	mov_avg
257	2011	9.578
258	2012	9.534
259	2013	9.570
260	2014	9.582
261	2015	9.608

[23]:
```
# Plot the two sets of data along with each other
fig, axes = plt.subplots(1, 2, figsize=(14, 4), squeeze=False)

hk_mov_avg_5.plot(x = "year", y = "mov_avg", ax = axes[0, 0], kind = "line",
```

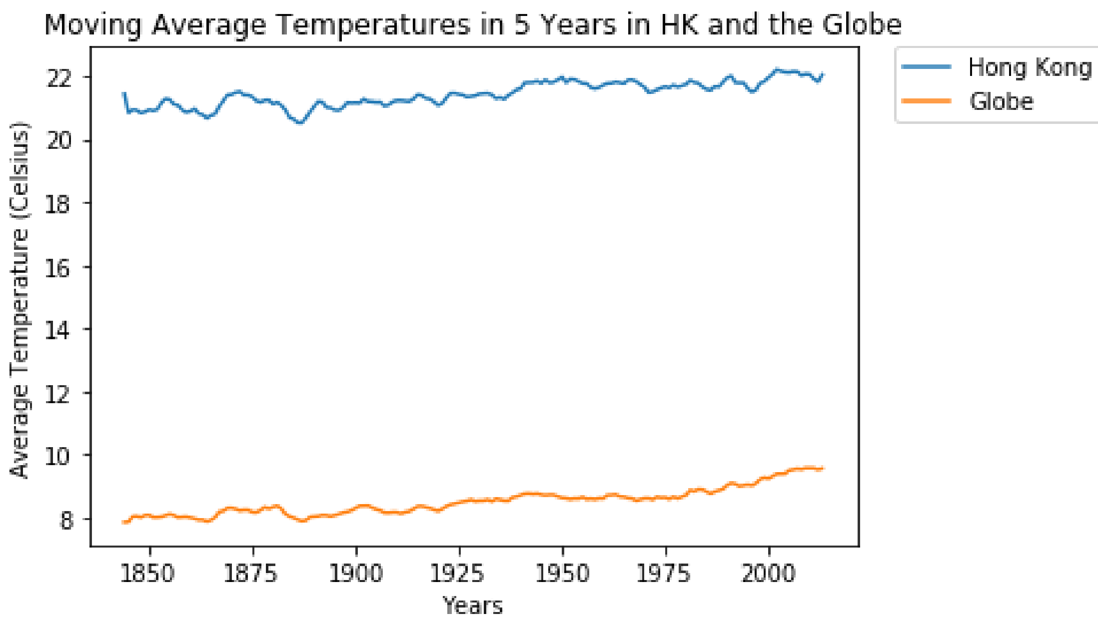
```
                    title = "Moving Average Temperatures in 5 years in HK")
global_mov_avg_5.plot(x = "year", y = "mov_avg", ax = axes[0, 1], kind = "line",
                        title = "Moving Average Temperatures in 5 years in the␣
 ↪Globe")
axes[0, 0].set_xlabel("Years")
axes[0, 0].set_ylabel("Average Temperature (Celsius)")
axes[0, 1].set_xlabel("Years")
axes[0, 1].set_ylabel("Average Temperature (Celsius)")
plt.show()
```



These two plots are line charts on the moving average temperatures in a five-year window for the globe and Hong Kong. Since both charts have different x and y axes, it would be hard to do a rough comparison between the globe and Hong Kong, and thus I plot them on the same scale with different colors in the following:

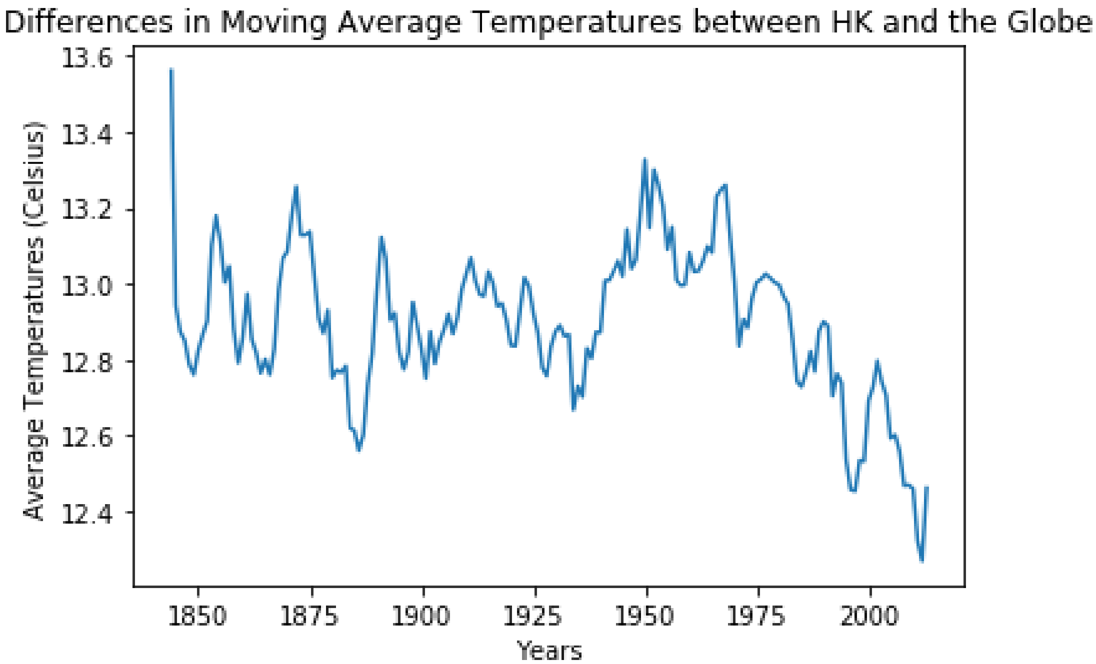
[24]:
```
# Plot the two sets of data on the same scale
# Obtain the global data subset with the same years of records as the HK data␣
 ↪set
global_mov_avg_5_subset = global_mov_avg_5[
    (global_mov_avg_5["year"] >= min(hk_mov_avg_5["year"])) &
    (global_mov_avg_5["year"] <= max(hk_mov_avg_5["year"]))]
global_mov_avg_5_subset.reset_index(inplace = True)

plt.subplot(111)
plt.plot(hk_mov_avg_5["year"], hk_mov_avg_5["mov_avg"], label="Hong Kong")
plt.plot(global_mov_avg_5_subset["year"], global_mov_avg_5_subset["mov_avg"],␣
 ↪label="Globe")
plt.title("Moving Average Temperatures in 5 Years in HK and the Globe")
plt.xlabel('Years')
plt.ylabel('Average Temperature (Celsius)')
# Place a legend to the right of this smaller subplot
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
plt.show()
```



To get a better idea about the changes in temperatures over time in Hong Kong as compared with those changes in the world, a line chart on the differences in moving average temperatures between the two during the years that are also present in both dataframe as follows:

[25]:
```
# Plot the differences in temperatures between HK and the globe
plt.plot(hk_mov_avg_5["year"],
         hk_mov_avg_5["mov_avg"] - global_mov_avg_5_subset["mov_avg"])
plt.title("Differences in Moving Average Temperatures between HK and the Globe")
plt.xlabel("Years")
plt.ylabel("Average Temperatures (Celsius)")
plt.show()
```



As we can see from the above chart, the differences in temperatures between Hong Kong and the world become smaller starting from around 1950. However, it does not mean that the temperatures in Hong Kong has become cooler, but such an interpretation issue will come up for further discussion in part IV.

Part III - Line Chart of Moving Average Temperatures (10-Year Window) In this part, I will plot the same charts as in part II, except that I will plot the moving average temperatures in a ten-year window instead of a five-year one, so that we can see if the smoothing effects are different given the different calculations of moving averages.

[26]:
```
# Obtain the moving average temperatures in 10-year window for hk_temp␣
↪dataframe
hk_temp_rolling_10 = hk_temp["avg_temp"].rolling(10).mean()
hk_mov_avg_year_10 = hk_temp["year"][hk_temp_rolling_10.dropna().index].
↪reset_index(drop=True)
hk_mov_avg_temp_10 = pd.Series(hk_temp_rolling_10.dropna().values).
↪rename("mov_avg")

#hk_mov_avg_year_10 = pd.Series([])
#hk_mov_avg_temp_10 = pd.Series([])
#for i in range(len(hk_temp["avg_temp"]) - 9):
#    hk_mov_avg_year_10[i] = hk_temp["year"][i+9] # Show the tenth year of␣
↪moving averages
#    hk_mov_avg_temp_10[i] = (hk_temp["avg_temp"][i:i+10]).mean() # Calculate␣
↪the moving averages
```

[27]:
```
# Combine the two Series of years and moving average temperatures into a␣
↪dataframe
hk_mov_avg_10 = pd.concat([hk_mov_avg_year_10, hk_mov_avg_temp_10], axis = 1)
hk_mov_avg_10.rename(columns = {0: "year", 1: "mov_avg"}, inplace = True) #␣
↪Rename the columns
hk_mov_avg_10.tail() # Use .tail() to see if it works
```

[27]:

	year	mov_avg
160	2009	22.081
161	2010	22.069
162	2011	22.016
163	2012	21.962
164	2013	22.021

[28]:
```
# Obtain the moving average temperatures in 10-year window for global_temp␣
↪dataframe
global_temp_rolling_10 = global_temp["avg_temp"].rolling(10).mean()
global_mov_avg_year_10 = global_temp["year"][global_temp_rolling_10.dropna().
↪index].reset_index(drop=True)
global_mov_avg_temp_10 = pd.Series(global_temp_rolling_10.dropna().values).
↪rename("mov_avg")

#global_mov_avg_year_10 = pd.Series([])
#global_mov_avg_temp_10 = pd.Series([])
#for i in range(len(global_temp["avg_temp"]) - 9):
#    global_mov_avg_year_10[i] = global_temp["year"][i+9] # Show the tenth year␣
↪of moving averages
#    global_mov_avg_temp_10[i] = (global_temp["avg_temp"][i:i+10]).mean() #␣
↪Calculate the moving averages
```

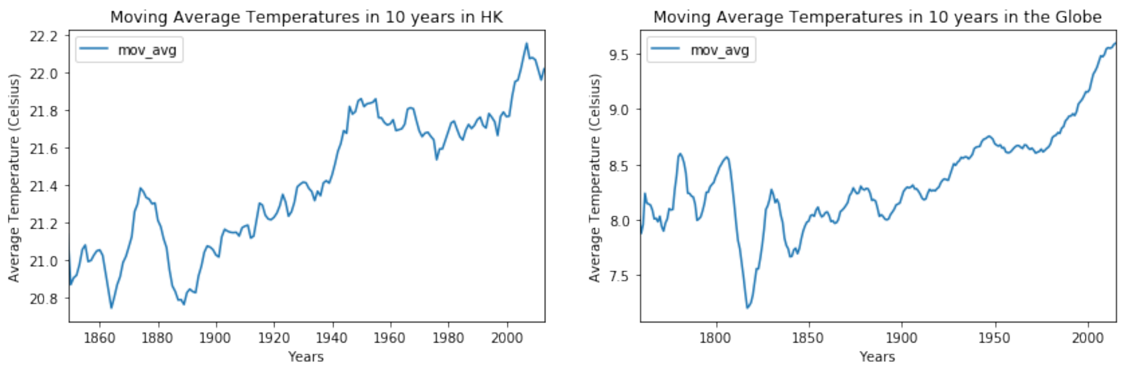
[29]:
```
# Combine the two Series of years and moving average temperatures into a␣
↪dataframe
global_mov_avg_10 = pd.concat([global_mov_avg_year_10, global_mov_avg_temp_10],␣
↪axis = 1)
global_mov_avg_10.rename(columns = {0: "year", 1: "mov_avg"}, inplace = True) #␣
↪Rename the columns
global_mov_avg_10.tail() # Use .tail() to see if it works
```

[29]:

	year	mov_avg
252	2011	9.554
253	2012	9.548
254	2013	9.556
255	2014	9.581
256	2015	9.594

[30]:
```
# Plot the two sets of data along with each other
fig, axes = plt.subplots(1, 2, figsize=(14, 4), squeeze=False)
hk_mov_avg_10.plot(x = "year", y = "mov_avg", ax = axes[0, 0], kind = "line",
                   title = "Moving Average Temperatures in 10 years in HK")
```

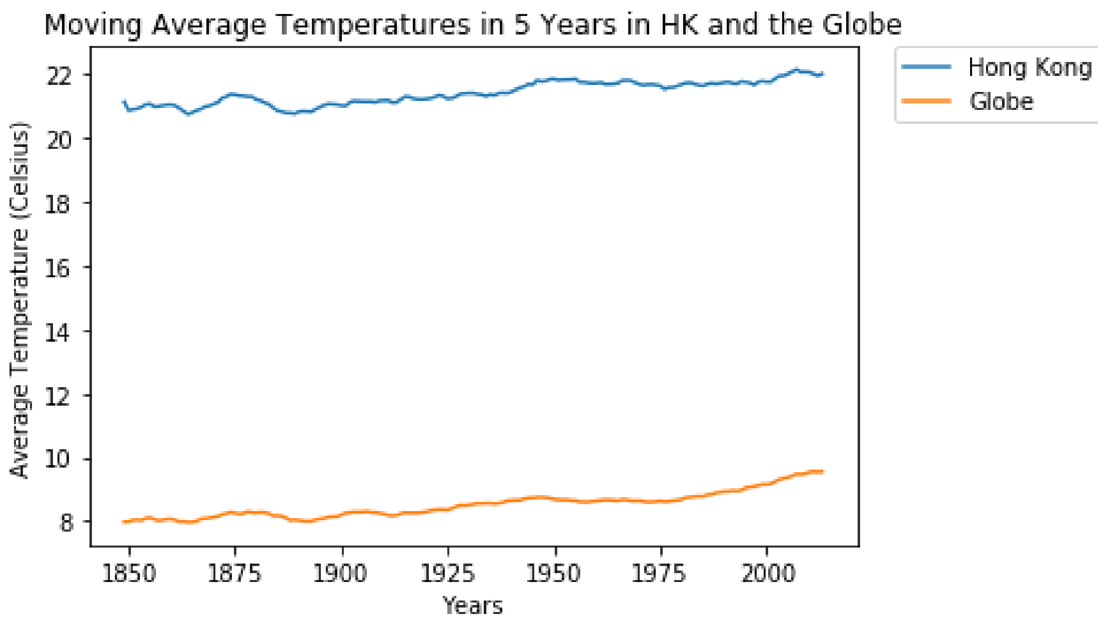
```
global_mov_avg_10.plot(x = "year", y = "mov_avg", ax = axes[0, 1], kind =␣
↪"line",
                        title = "Moving Average Temperatures in 10 years in the␣
↪Globe")
axes[0, 0].set_xlabel("Years")
axes[0, 0].set_ylabel("Average Temperature (Celsius)")
axes[0, 1].set_xlabel("Years")
axes[0, 1].set_ylabel("Average Temperature (Celsius)")
plt.show()
```



These two plots are line charts on the moving average temperatures in a ten-year window for the globe and Hong Kong. Since both charts have different x and y axes, it would be hard to do a rough comparison between the globe and Hong Kong, and thus I plot them on the same scale with different colors in the following:

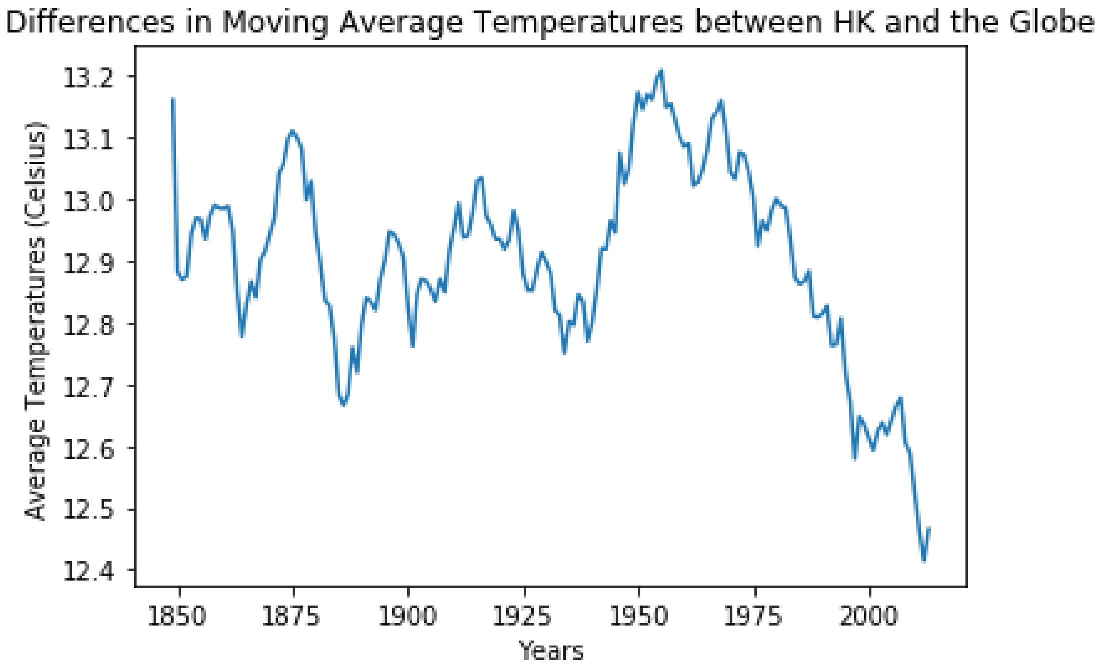
[31]:
```
# Plot the two sets of data on the same scale
# Obtain the global data subset with the same years of records as the HK data␣
↪set
global_mov_avg_10_subset = global_mov_avg_10[
    (global_mov_avg_10["year"] >= min(hk_mov_avg_10["year"])) &
    (global_mov_avg_10["year"] <= max(hk_mov_avg_10["year"]))]
global_mov_avg_10_subset.reset_index(inplace = True)

plt.subplot(111)
plt.plot(hk_mov_avg_10["year"], hk_mov_avg_10["mov_avg"], label="Hong Kong")
plt.plot(global_mov_avg_10_subset["year"], global_mov_avg_10_subset["mov_avg"],␣
↪label="Globe")
plt.title("Moving Average Temperatures in 5 Years in HK and the Globe")
plt.xlabel('Years')
plt.ylabel('Average Temperature (Celsius)')
# Place a legend to the right of this smaller subplot
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
plt.show()
```



To have a better idea about the changes in temperatures over time in Hong Kong as compared with those changes in the world, a line chart on the differences in moving average temperatures between the two during the years that are also present in both dataframe as follows:

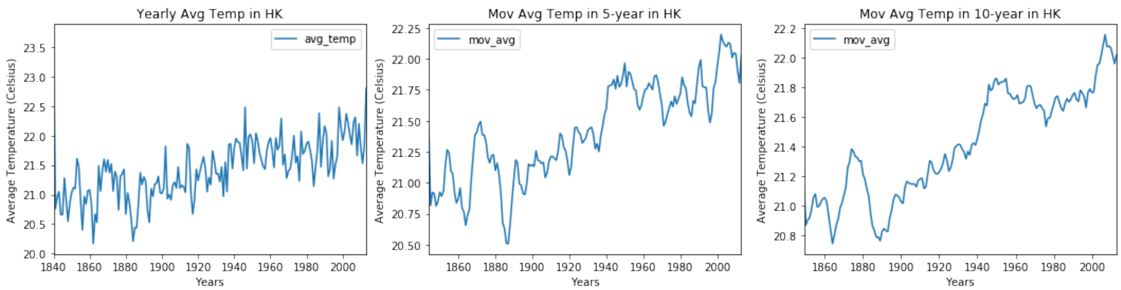
[32]:
```
# Plot the difference in temperatures between HK and the globe
plt.plot(hk_mov_avg_10["year"],
         hk_mov_avg_10["mov_avg"] - global_mov_avg_10_subset["mov_avg"])
plt.title("Differences in Moving Average Temperatures between HK and the Globe")
plt.xlabel("Years")
plt.ylabel("Average Temperatures (Celsius)")
plt.show()
```



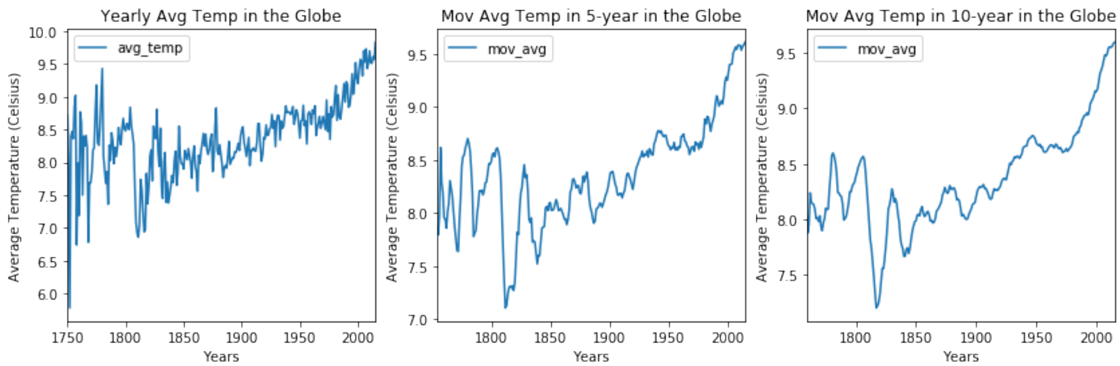
As we can see from the above chart, the differences in temperatures between Hong Kong and the world started to decrease at around 50s, and the trend appears to be more observable here than the previous chart using moving average in a 5-year window. Let's head to the next part wih a detailed discussion on observations and interpretations from the above line charts.

Part IV - Observations on the Temperature Trends from the Line Charts For the sakes of highlighting the smoothing effects and minimizing scrolling, the above plots will be shown again together with its counterpart versions in the following:

[33]:
```
# Plot the yearly average and moving average temerpatures in HK
fig, axes = plt.subplots(1, 3, figsize=(18, 4), squeeze=False)
hk_temp.plot(x = "year", y = "avg_temp", ax = axes[0, 0], kind = "line",
             title = "Yearly Avg Temp in HK")
hk_mov_avg_5.plot(x = "year", y = "mov_avg", ax = axes[0, 1], kind = "line",
                  title = "Mov Avg Temp in 5-year in HK")
hk_mov_avg_10.plot(x = "year", y = "mov_avg", ax = axes[0, 2], kind = "line",
                   title = "Mov Avg Temp in 10-year in HK")
axes[0, 0].set_xlabel("Years")
axes[0, 0].set_ylabel("Average Temperature (Celsius)")
axes[0, 1].set_xlabel("Years")
axes[0, 1].set_ylabel("Average Temperature (Celsius)")
axes[0, 2].set_xlabel("Years")
axes[0, 2].set_ylabel("Average Temperature (Celsius)")
plt.show()
```

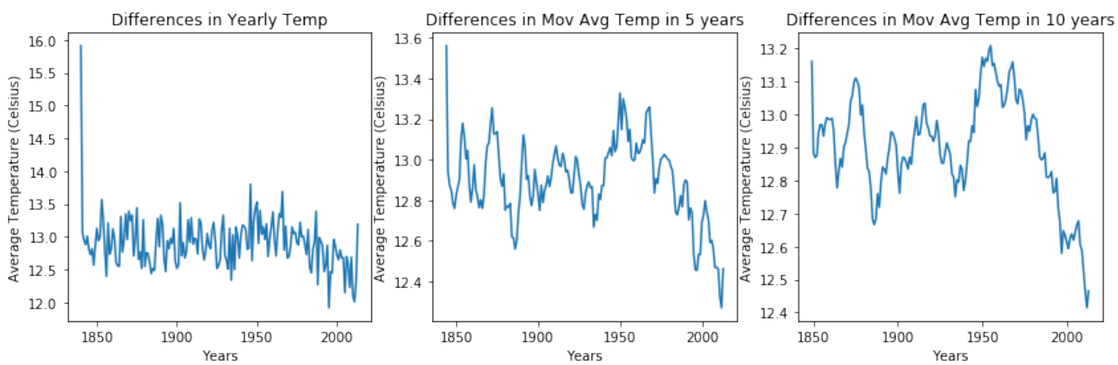


[34]:
```
# Plot the yearly average and moving average temerpatures in the globe
fig, axes = plt.subplots(1, 3, figsize=(14, 4), squeeze=False)
global_temp.plot(x = "year", y = "avg_temp", ax = axes[0, 0], kind = "line",
                 title = "Yearly Avg Temp in the Globe")
global_mov_avg_5.plot(x = "year", y = "mov_avg", ax = axes[0, 1], kind = "line",
                      title = "Mov Avg Temp in 5-year in the Globe")
global_mov_avg_10.plot(x = "year", y = "mov_avg", ax = axes[0, 2], kind =␣
 ↪"line",
                       title = "Mov Avg Temp in 10-year in the Globe")
axes[0, 0].set_xlabel("Years")
axes[0, 0].set_ylabel("Average Temperature (Celsius)")
axes[0, 1].set_xlabel("Years")
axes[0, 1].set_ylabel("Average Temperature (Celsius)")
axes[0, 2].set_xlabel("Years")
axes[0, 2].set_ylabel("Average Temperature (Celsius)")
plt.show()
```



[35]:
```
fig, axes = plt.subplots(1, 3, figsize=(14, 4), squeeze=False)
plt.sca(ax = axes[0, 0])
plt.plot(hk_temp["year"],
         hk_temp["avg_temp"] - global_temp_subset["avg_temp"])
plt.title("Differences in Yearly Temp")
plt.sca(ax = axes[0, 1])
```

```
plt.plot(hk_mov_avg_5["year"],
         hk_mov_avg_5["mov_avg"] - global_mov_avg_5_subset["mov_avg"])
plt.title("Differences in Mov Avg Temp in 5 years")
plt.sca(ax = axes[0, 2])
plt.plot(hk_mov_avg_10["year"],
         hk_mov_avg_10["mov_avg"] - global_mov_avg_10_subset["mov_avg"])
plt.title("Differences in Mov Avg Temp in 10 years")
axes[0, 0].set_xlabel("Years")
axes[0, 0].set_ylabel("Average Temperature (Celsius)")
axes[0, 1].set_xlabel("Years")
axes[0, 1].set_ylabel("Average Temperature (Celsius)")
axes[0, 2].set_xlabel("Years")
axes[0, 2].set_ylabel("Average Temperature (Celsius)")
plt.show()
```



Generally speaking, the temperatures in both Hong Kong and the globe appear to increase in the overall trend, so both Hong Kong and the world are getting hotter. The temperature trend in Hong Kong has been rising since 1900, but the temperature has even jumped higher than the increasing trend from around 2000; on the other hand, the temperature trend in the globe has been rising since 1850, but the temperature has rocketed from around 1970. Thus, the trend has not been consistent over the last few hundred years, in such a way that both Hong Kong and the world are getting much hotter than expected as well. In other words, the temperatures in both Hong Kong and the globe increase in the overall trend, and the temperatures increase much faster than the expected increasing trend.

Despite the similar trends between Hong Kong and the world, the trend of the differences in temperatures between them has been decreasing since 1950, so the difference has not been consistent over time. Such a decrease in the temperature difference indicates that Hong Kong is getting cooler on average compared to the global average, but it is mainly led by the increase in the global temperature. With reference to the line chart on the moving average temperatures (both in the 5-year and 10-year window) in the world, the huge rise from around 1970 to around 2010 is almost 1 degree, which is large given the mean temperature is 8.37°. In contrast, the moving average temperatures (both in the 5-year and 10-year window) in Hong Kong were fluctuated between 21.6° and 21.8° from 1950 to 2000, and had started to increase from 21.8° to 22.2° only since 2000, where such an increase of 0.4 degree is small given the mean temperature is 21.43°. That is to say, the temperature in Hong Kong has increased steadily overall while the temperature in the globe has increased suddenly since 1970, and the world temperature has increased much faster than Hong Kong temperature, i.e. the globe is getter hotter much faster than Hong Kong.

Conclusion

This data analysis is divided into two main parts: data wrangling and data visualization. In data wrangling, I gathered the data by using SQL queries to extract the data from Udacity's database, and assessed and cleaned the data by using Pandas. In the data visualization, I created line charts by using Matplotlib and calculating the moving averages in 5-year and 10-year windows to make the temperature trends more observable. Apart from making trends more observable, I tried to plot both the temperature trends in the world and Hong Kong in the same chart in order to make a direct comparison between them, but the visual fails to help me make a comparison given the huge difference between the global and local mean temperatures as well as different initial years of record, so I decided to plot the difference in temperatures in both the globe and Hong Kong.

On the basis of the data visualization, it is observed that (1) the temperatures in both Hong Kong and the globe increase in the overall trend, and (2) the global and local temperatures increase much faster than the expected increasing trend, i.e. the world and my city are getting hotter than what has been expected from previous data. There are, however, differences in the temperature trends between the world and Hong Kong. Given the decreasing trend in the difference between the local and global temperature, we might say that (3) Hong Kong is cooler on average compared to the global average, and this is obviously brought by the greater increase in global temperature than the local counterpart. Moreover, it is discovered that (4) the temperature in Hong Kong has increased steadily overall while the temperature in the globe has increased suddenly since 1970, and (5) the world temperature has increased much faster than Hong Kong temperature, i.e. the globe is getter hotter much faster than Hong Kong.

Reference

Wickham, Hadley. "Tidy Data". In The Journal of Statistical Software. Volume 59, 2014.

[ ]:

```

```