Data Analyst Nanodegree Assignment 1: Exploring Weather Trends

Huey Ning Lok

September 30th, 2018.

Project Goal

Analyze local and global temperature data and compare the temperature trends between the two.

Procedure

In preparation for visualization and analysis, selected portions of the raw data were extracted into a csv file using SQL.

SQL Code

```
SELECT cd.city, cd.year, cd.avg_temp AS city_avg_temp, gd.avg_temp AS
global_avg_temp
FROM city_data cd, global_data gd
WHERE cd.year=gd.year AND cd.city='Kuala Lumpur'
```

The csv file was then imported into Python for data manipulation, and to create the appropriate graphs. A NaN count revealed that the 'city_avg_temp' column had 17 missing values (out of 189 total rows). These empty cells were filled with the value from the preceding cell, i.e. the average city temperature of the year before, under the presumption that:

- The average temperature of a city in any given year should not differ too greatly from the temperature the year before.
- Even a worst-case scenario of 17 sequentially missing-value rows should not significantly impact our high-level analysis of the temperature trend over 189 years.

The pandas library was heavily relied on for easy manipulation of the data, including calculating the moving average using the `'pandas.Dataframe.rolling'` function. The moving window was set at 10 years, and the minimum number of observations required to have a valid value was set at 1.

Python Code

```
#import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
sns.set_context("notebook", font_scale=1.5, rc={"lines.linewidth": 2.5})

#load the data
avg_temp = pd.read_csv('avg_temp.csv')

#count number of NaN values in data
nan_count = avg_temp.isna().sum()
print('nan_count: ',nan_count)

#Replace all NaN values for city_avg_temp with value in preceding cell
avg_temp['city_avg_temp'].fillna(method='ffill',inplace=True)

#calculate 10-year Moving Avg for city and global temps
avg_temp['City temp 10-year
MA']=avg_temp['city_avg_temp'].rolling(10,min_periods=1).mean()
avg_temp['Global temp 10-year
MA']=avg_temp['global_avg_temp'].rolling(10,min_periods=1).mean()

#plot KL temp 10-year MA over the years
plt.figure(figsize=(20, 10))
plt.title("Kuala Lumpur Temperature Over the Years",fontsize=18)
plt.xlabel("Year",fontsize=18)
plt.ylabel("Temperature",fontsize=18)
plt.plot(avg_temp['year'],avg_temp['City temp 10-year MA'])

#plot Global temp 10-year MA over the years
plt.figure(figsize=(20, 10))
plt.title("Global Temperature Over the Years",fontsize=18)
plt.xlabel("Year",fontsize=18)
plt.ylabel("Temperature",fontsize=18)
plt.plot(avg_temp['year'],avg_temp['Global temp 10-year MA'])
```

Data Visualization

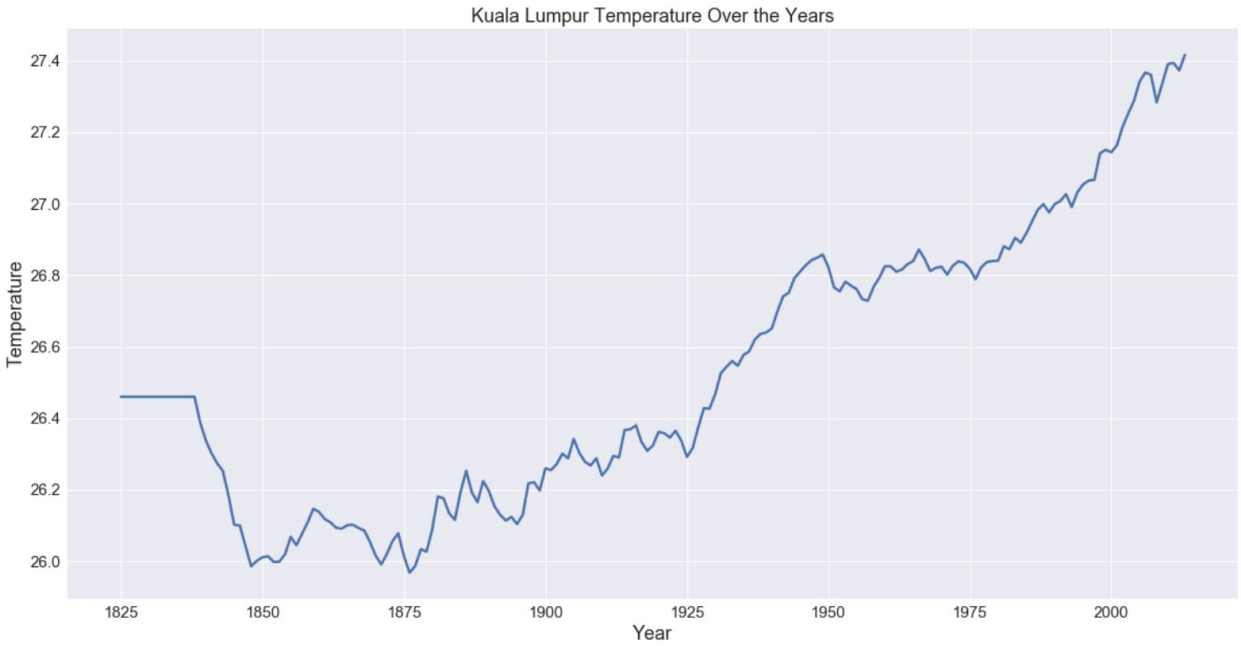

Figure 1

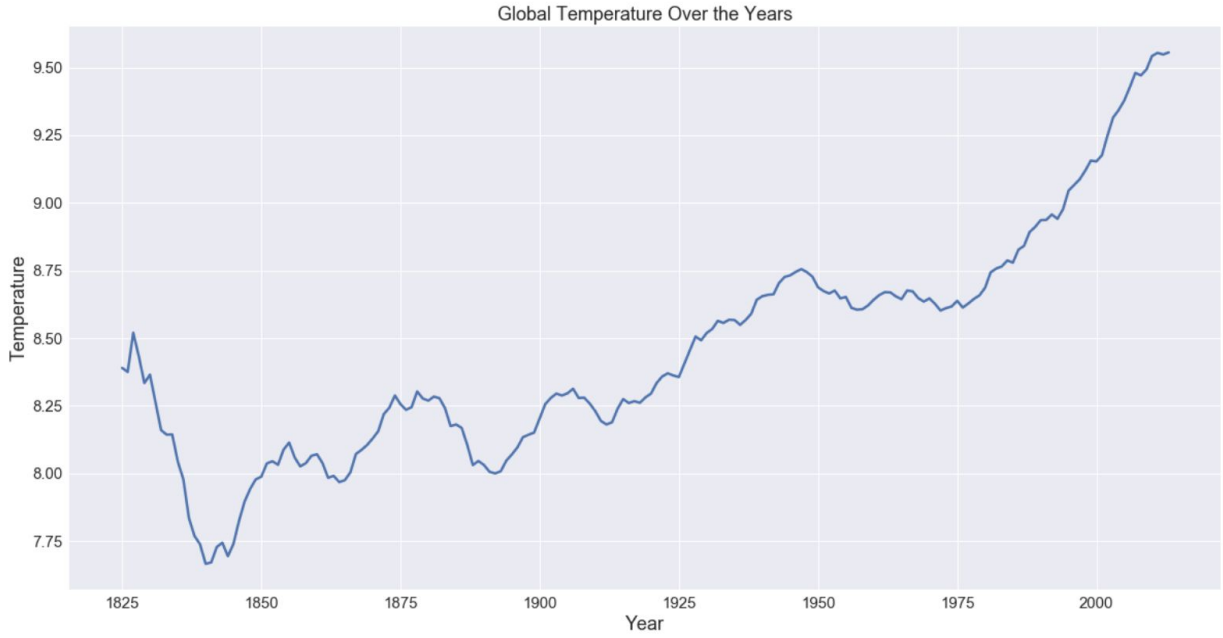

Figure 2

Data Analysis

Figure 1 displays the 10-year moving average of Kuala Lumpur's annual temperature from the year 1825 - 2013, while Figure 2 displays the 10-year moving average of global annual temperature over the same time period. It should be noted that both graphs have different y-axes scales: Figure 1 has a y-axis range of approximately 26 - 27 degree celsius; Figure 2 has a y-axis range of approximately 7.5 - 9.5 degree celsius. This difference in y-axes immediately tells us that Kuala Lumpur has a much higher average temperature compared to the global temperature, which makes sense considering it is a temperate country.

Although Kuala Lumpur has a much higher average temperature vs the global average, the two graphs exhibit a similar trend in temperature change. We see a decrease in temperature from the year 1825 to the year 1840 - where a global minima is located in both graphs. This is followed by a gradual increase towards the current maxima in the most recent years. The difference in temperature between Kuala Lumpur and the global average are very consistent - with the peaks and drops almost lining up exactly from one graph to another. For example, when global temperature was at its lowest of ~7.5°C around the year 1840, Kuala Lumpur was also at its lowest of ~26°C. At the global highest of ~9.5°C in 2013, Kuala Lumpur is at ~27.5°C. From 1840 - 2013, Kuala Lumpur's temperature increased by ~1.5°C, while global temperature increased by ~2°C.

The positive correlation between temperature and years signifies that despite little drops and peaks throughout the years, our global temperature is on a general upwards trend, and will most likely continue climbing. It is particularly worrying that the steepest incline, aka fastest rate of change in temperature, seems to have occured in the last 30 years or so. If the upwards trend continues, along with an ever increasing rate of change, the threats of global warming will only increase in prominence.