Data Analysis Project:
Yearly Average Temperatures

By Sam Celarek

Retrieving the Data

To retrieve the data for this project, three SQL queries were used:

```sql
SELECT year as "Year", city, avg_temp as "City Temp"
FROM city_data;
```

```sql
SELECT DISTINCT city
FROM city_data
ORDER BY city ASC
```

```sql
SELECT year as "Year", avg_temp as "Global Temp"
FROM global_data;
```

The first query retrieved the city data and reformatted it slightly, the second query retrieved the global data and reformatted it, and the third query retrieved all the unique instances of cities in the city data. The third query was not used in the actual data analysis, but was used as a reference when plugging in different cities into the analysis.

Computing the Moving Average

For this analysis, the pandas, numpy, and matplotlib libraries in Python were used, within the Visual Studio Code IDE in a IPython Notebook. The following is some starter code:

```python
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd

city_wth = pd.read_csv('All City.csv')
glob_wth = pd.read_csv('Global.csv')

# Selecting a city, and cleaning that city's temperature data
city_name = 'Portland'
port =
city_wth[city_wth['city']==f'{city_name}'].reset_index(drop=True
).copy()

port[f'{city_name} Temp'] = port['City Temp']
port.drop(['City Temp', 'city'], axis = 1, inplace=True)

# Then cleaning up the global data to match my city temp data
wth_merge = pd.merge(port, glob_wth, on='Year', how='inner')
glob = wth_merge[['Year', 'Global Temp']].copy()
```

To compute a moving average in Python, the pandas library function `.rolling()` was used (https://www.geeksforgeeks.org/how-to-calculate-moving-average-in-a-pandas-dataframe/). A 5-year window was used for the moving average because this is more standard in climate literature (https://rdcu.be/c2Spn). The following code was:

```
port[f'{city_name} Temp Moving Average'] = port[f'{city_name} Temp'].dropna().rolling(5).mean()

glob['Global Temp Moving Average'] = glob['Global Temp'].dropna().rolling(5).mean()
```

Unfortunately, if there was a `Nan` value in the moving average window, the function would output a `Nan`. The `.dropna()` function on both the datasets before computing the moving average.

Computing the Line of Best Fit

While the moving averages plots helped give an intuition for the trend of each dataset, the Least Squares Regression Line aka a Line of Best Fit illustrated trend of the data even better. The numpy functions `.polyfit()`, and `.poly1d()` were used to accomplish this. Here is the code used:

```
def line_of_best_fit(x, y, dataframe):

    # Fit a polynomial to the data
    data = pd.DataFrame({'x': x, 'y': y}).dropna()

    # Returns the coefficients for y = mx + b
    coefficients = np.polyfit(data['x'], data['y'], 1)

    # plugs these coefficients into the equation y = mx + b
    polynomial = np.poly1d(coefficients)

    # then I use this equation to do a linear transformation on the array of x
    # this new series of data are the points on my line of best fit
    dataframe['Line Of Best Fit'] = polynomial(x)

    return polynomial # this will return the slope and y-intersect which became important later.

port_poly = line_of_best_fit(port['Year'], port[f'{city_name} Temp'], port)
glob_poly = line_of_best_fit(glob['Year'], glob['Global Temp'], glob)

final = pd.merge(glob, port, on="Year", how='outer', suffixes=[' Global', f' {city_name}'])

```

Visualizations and Observations

Portland, where I live, happened to be in the `city_data` database and so I compared it to the global temperature trends. These are the first 5 insightful visualizations I made for this data.

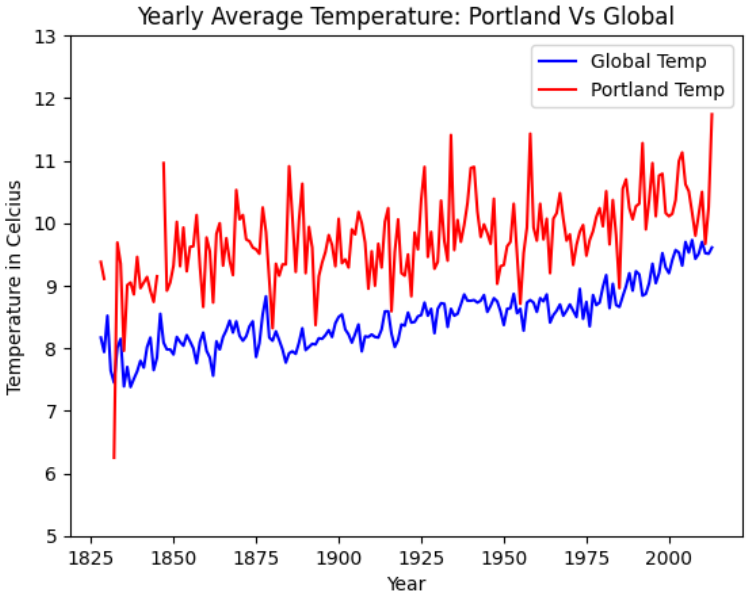

Figure 1

Figure 1 is the unaltered data from Portland's yearly average temperature plotted against its global counterpart over time in years. This visualization is noisy with lots of variance, but it illustrates that Portland's yearly average temperature is warmer than the global temperature and potentially linearly increasing. This also helps us ground the expectations for what we should see in later, more smoothed out plots.

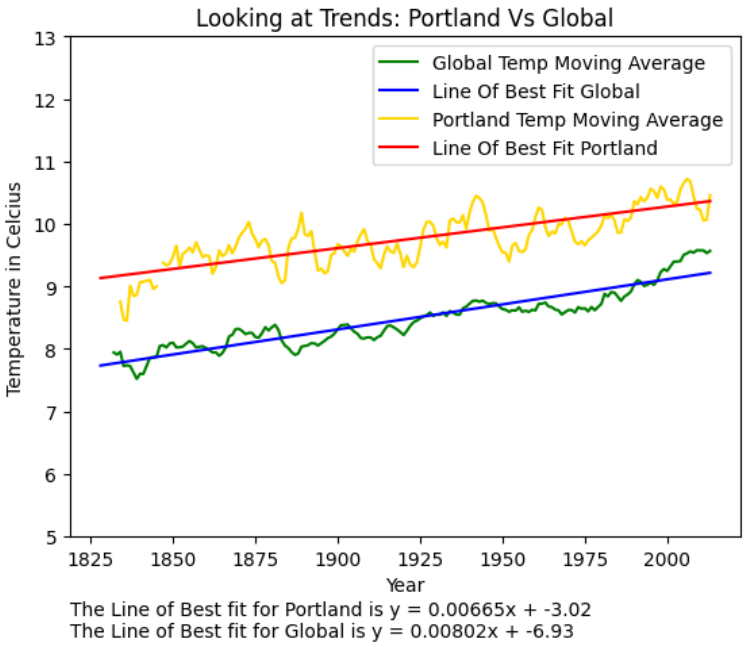

Figure 2

Figure 2 uses both the moving average with a 5-year window and the line of best fit for both datasets to better demonstrate the trend of each. This makes it very intuitively obvious that both Portland and global temperatures have been rising consistently since 1825 in our dataset. However it is not obviously which temperature is increasing faster.

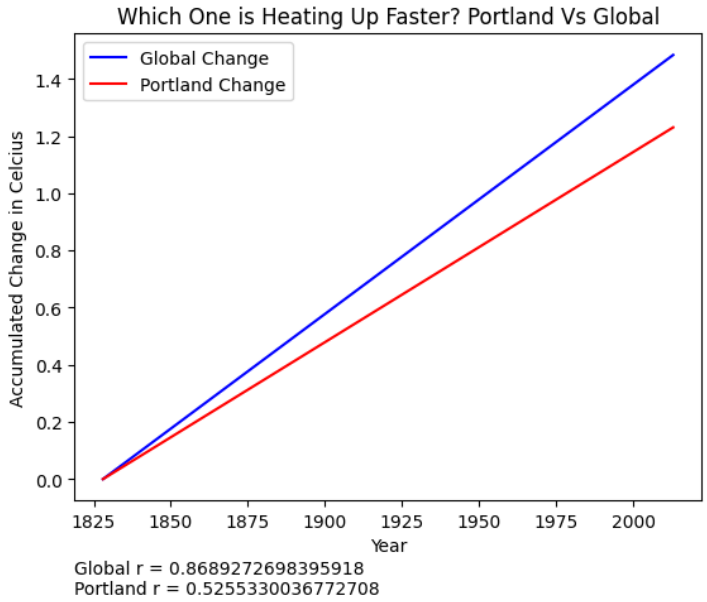

Figure 3

Figure 3 illustrates clearly that global temperature yearly averages predicted by the line of best fit are increasing faster than Portland's. While lines of best fit are just approximations and only work for linear trends, both lines of best fit had r values above 0.5, which roughly means that given some year, the predictive accuracy of that year's temperature is moderate to high. So the next question is, if Portland started out warmer but is now warming slower than its global counterpart, then how much of a difference in temperatures remain?

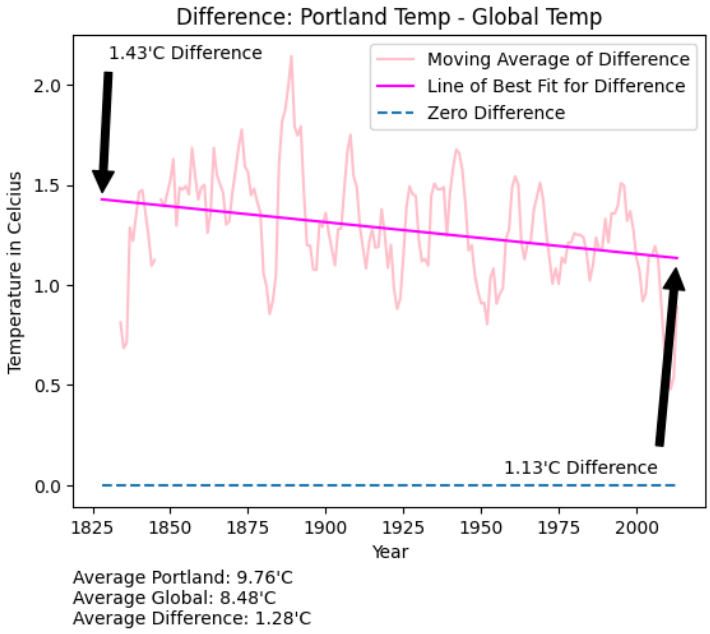

Figure 4

Figure 4 is a graph of Portland yearly average temperatures minus global temperatures, displayed as a moving average with a 5-year window and with a line of best fit. The trend in this graph is very subtle and has lots of variance (even when just looking at the moving average), but it would appear that the difference between Portland and global temperatures is slowly shrinking from an expected 1.43'C each year to an expected 1.13'C. To irresponsibly extrapolate beyond our dataset just for fun, this would imply that the Portland and global yearly temperature averages will converge in 2728. This is super wrong, but interesting nonetheless!

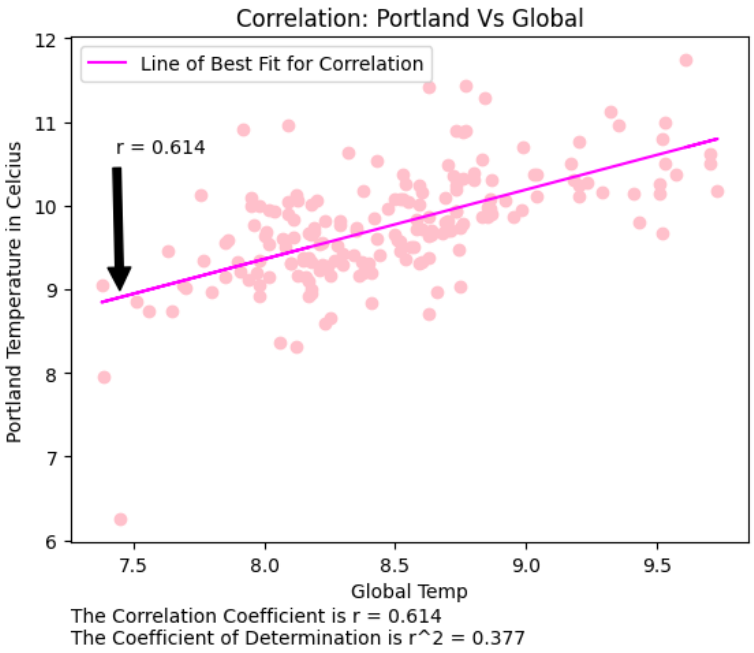

Figure 5

In Figure 5, a scatter plot with a line of best fit was used to illustrate the relationship between Portland and global temperatures. The correlation between the two is moderately positive, with an r value of 0.614 and an r^2 value of 0.377, indicating that 37.7% of the variance in Portland temperatures can be explained by global temperatures.

The 1820 Inflection Point

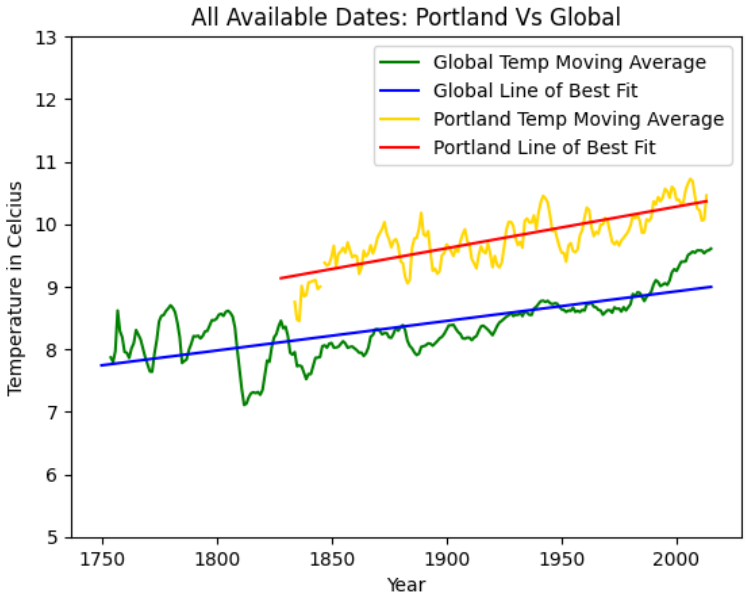

Figure 6

Interestingly, I noticed that the upwards trend in global temperatures is a recent phenomenon in the data that starts around 1835. This wasn't as apparent in the first visualizations because the Portland data starts in 1825 and thus the comparison between global and city temperatures began in 1825.

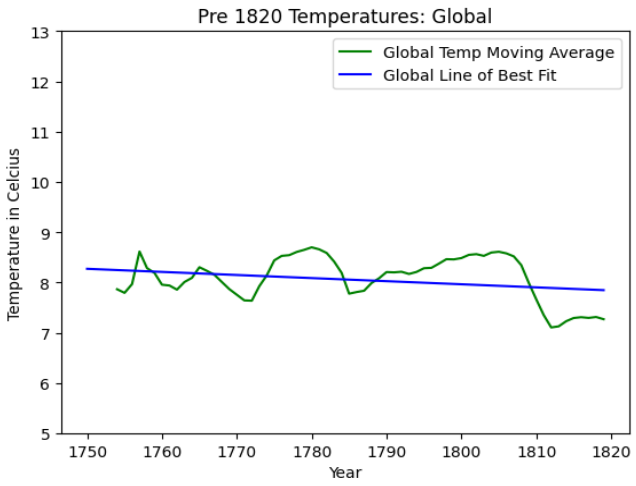

Figure 8

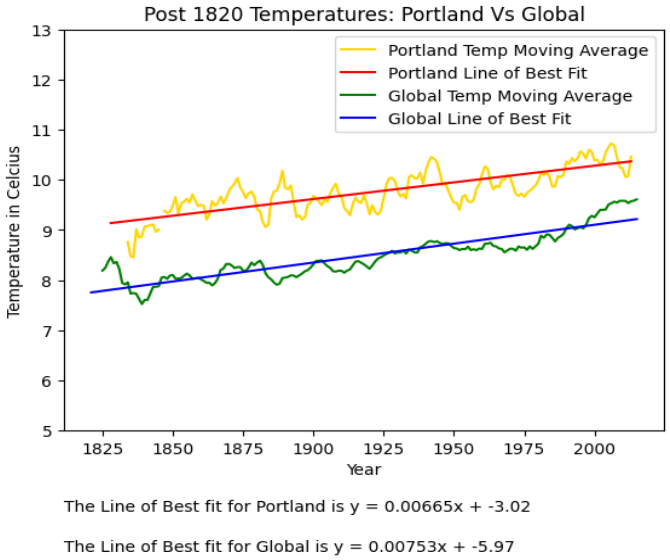

Figure 7

Figure 7 and 8 above display global temperature trends before and after 1820.

Figure 9 below captures how different the global trends have been before and after 1820. Why is 1820 such an inflection point in global temperature data?

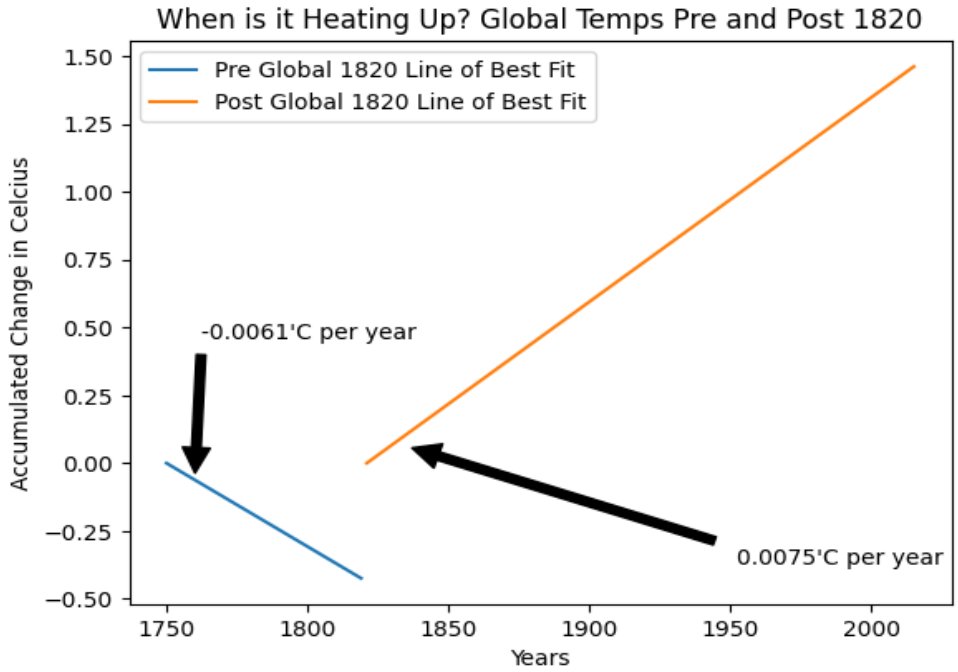

Figure 9

One possible explanation could be the start of the industrial revolution and increased release of CO2 through burning coal, but CO2 levels didn't rise until 1900 (https://www.statista.com/statistics/264699/worldwide-co2-emissions/). Another explanation is maybe there are natural climate cycles and before 1820 we were in the downswing of a cycle and since then we are on the upswing, but there is not much evidence to support that either (https://wires.onlinelibrary.wiley.com/share/JZGJ3CP6Y2AFW4BSUCVQ?target=10.1002/wcc.18 , https://royalsociety.org/topics-policy/projects/climate-change-evidence-causes/question-6/).

However, there is another more banal explanation: measurement error and sampling bias. The methods for collecting 'global' temperatures were not as precise nor global before 1820. For example, collecting the temperature of the North and South Poles is logistically very hard to obtain and would have *pulled* the global temperature in the pre-1820 period towards a colder average. Further supporting this interpretation, by 1820 only 44.7% (147/329) of cities were reporting temperature data. This number improved rapidly thought to 98.8% (325/329) by 1860, and 100% by 1882.

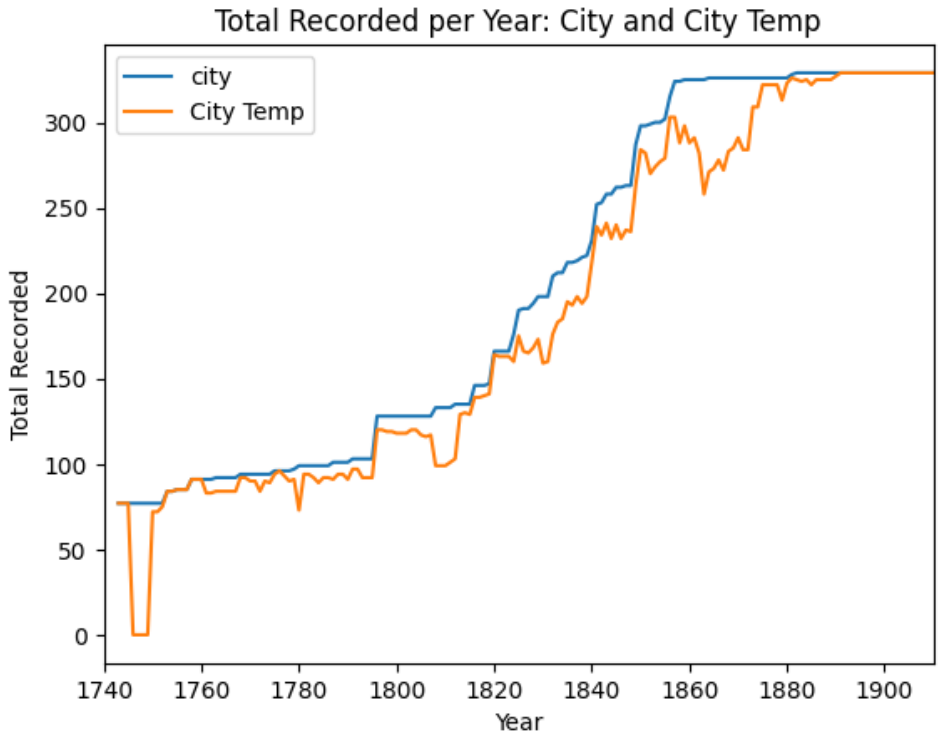

Figure 10

Extra Visualizations

These are simply some other city visualizations that I was able to make with my program.

Seoul:

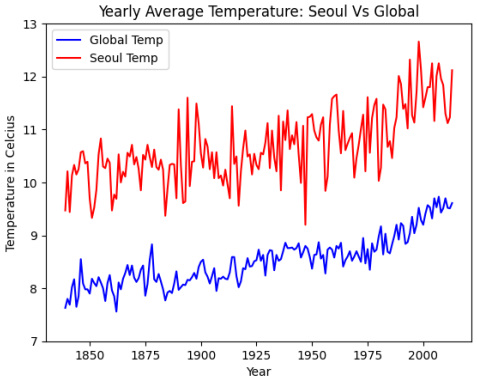

Figure 13

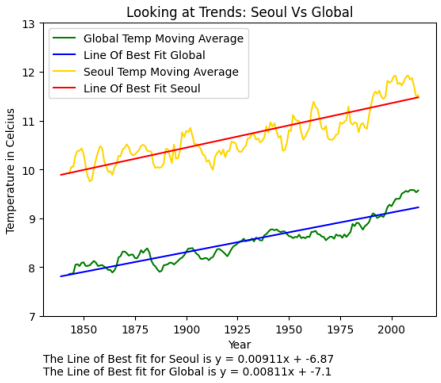

Figure 12

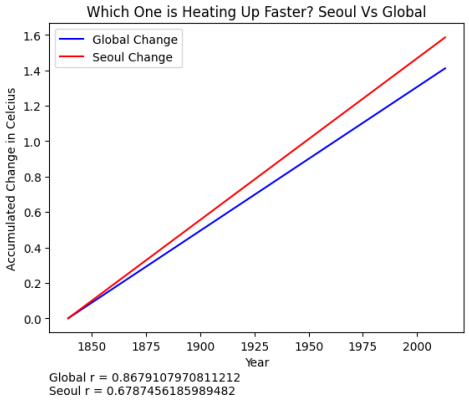

Figure 11

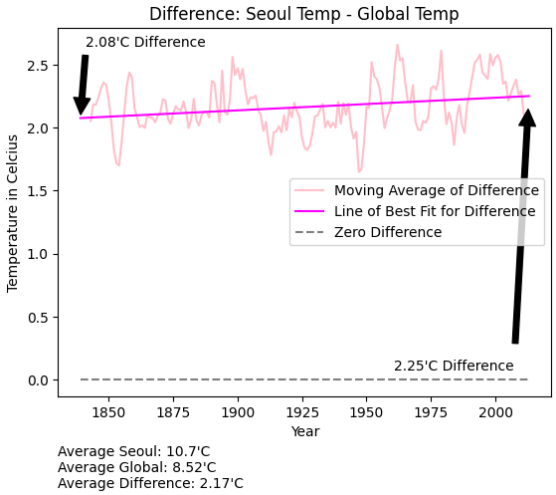

Figure 15

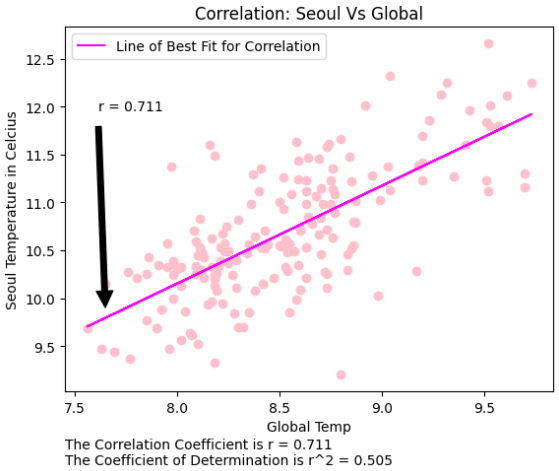

Figure 14

Observations: From Figures 11-15 we can tell that Seoul is much warmer than the global averages, is warming faster than the globe, and is very slightly diverging from the global temperature more as time goes on. Seoul's temperature correlates strongly with the global temperature with r = 0.711.

San Francisco:

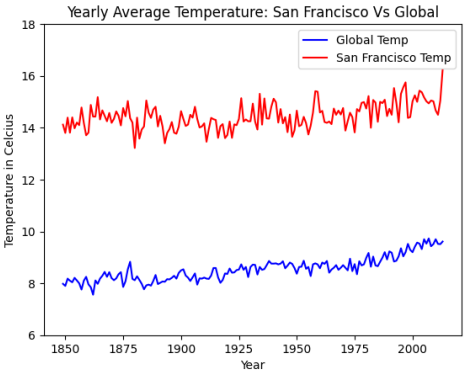

Figure 18

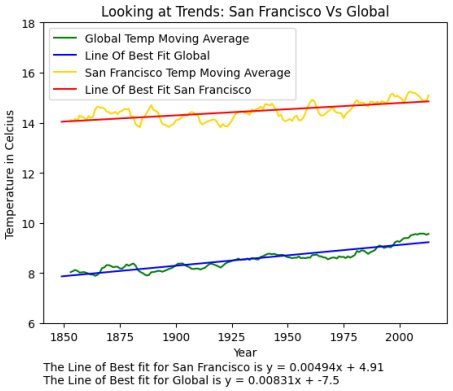

Figure 17

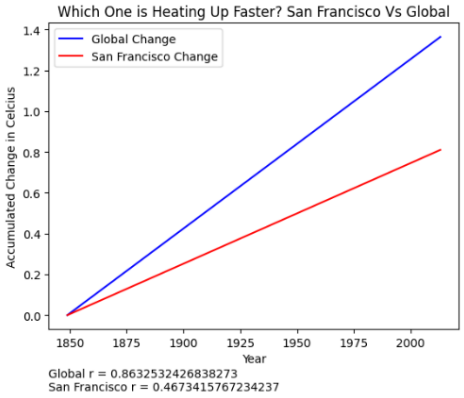

Figure 16

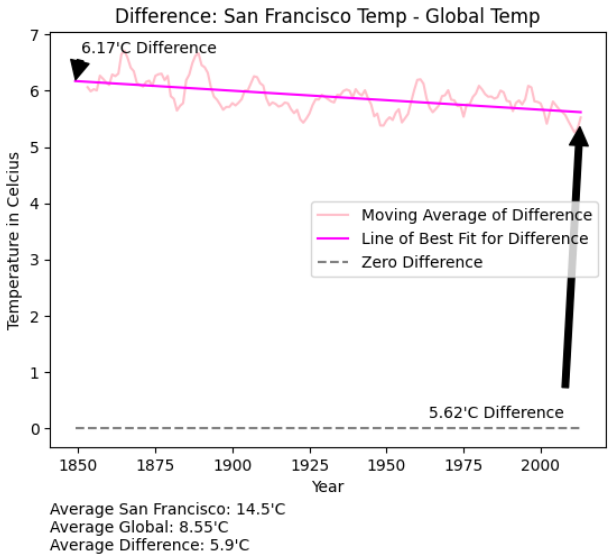

Figure 20

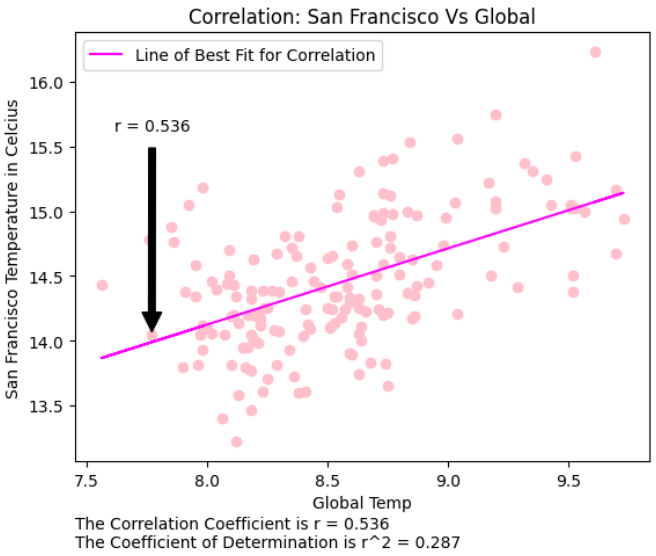

Figure 19

Observations: Figures 16-20 reveal that San Francisco is much warmer than the globe, but it is warming slower, and thus the gap between San Francisco's average yearly temperature and the globes is decreasing. San Francisco's temperature correlates the least with the global temperature of all the cities with an r value of 0.536, although this would still be considered moderately positive.