Kalman Heyn
January 8, 2020

Udacity Project 1: Exploring Weather Trends

Introduction

A data analyst is responsible for gathering, manipulating, and deciphering data. This first project was to exercise the basic skills of a data analyst and apply them to real world data – in this case, global and local weather trends.

The following report includes an outline of the tools used in gathering the data, the calculations involved in its manipulation, and key considerations in visualizing its trends. These visualizations are included below and accompanied by relevant observations.

Outline

What tools were used for each step?

1. **SQL** – The first step was to pull the data from the temperatures database. The data was extracted into CSV files. The following commands were run to get data from the city_list, city_data, and global_data tables:

   ```
   SELECT *
   FROM global_data;
   ```

   ```
   SELECT *
   FROM city_list;
   ```

   ```
   SELECT *
   FROM city_data
   WHERE city = 'San Francisco';
   ```

2. **Microsoft Excel** – The data was pooled in Excel where the moving average was calculated.
3. **Microsoft Excel Charts** – After moving averages were calculated, the data was displayed in multiple charts within Microsoft Excel.

How did you calculate the moving average?

Five averages were calculated to explore the effects of a longer or shorter moving average:

- Chart 1 displays the data as it was provided.
- Chart 2 displays the data with a five-year moving average.
- Chart 3 displays the data with a ten-year moving average.
- Chart 4 displays the data with a twenty-year moving average.
- Chart 5 displays the data with a fifty-year moving average.

The general formula for moving averages is as follows:

A *n*-year moving average is calculated by adding the current year, y_c, to the previous year, y_{c-1}, and so on until there are *n* number of years being added. The sum is then divided by *n* to find the *n*-year moving average:

$$n\text{-year moving average} = \frac{y_c + y_{c-1} + \cdots + y_{c-(1-n)}}{n}$$

For example, calculating a 5-year moving average for global weather in 2015 is as follows:

$$5\text{-year moving average} = \frac{y_{2015} + y_{2015-1} + y_{2015-2} + y_{2015-3} + y_{2015-4}}{5}$$

$$5\text{-year moving average} = \frac{9.83 + 9.57 + 9.61 + 9.51 + 9.52}{5}$$

$$5\text{-year moving average} = 9.608$$

In Microsoft Excel, this is just the formula =AVERAGE(y_c:$y_{c-(n-1)}$). It is called a moving average because this process is done for every point in the dataset except for the first *n* points.

What were your key considerations when deciding how to visualize the trends?

- How can I visually differentiate between global and San Francisco trends?
- What bounds should I use on the Y-axis to capture the trend characteristics?
- What bounds should I use on the X-axis to capture the most important data points?
- How many years should I capture in the moving average?

Charts

Chart 1

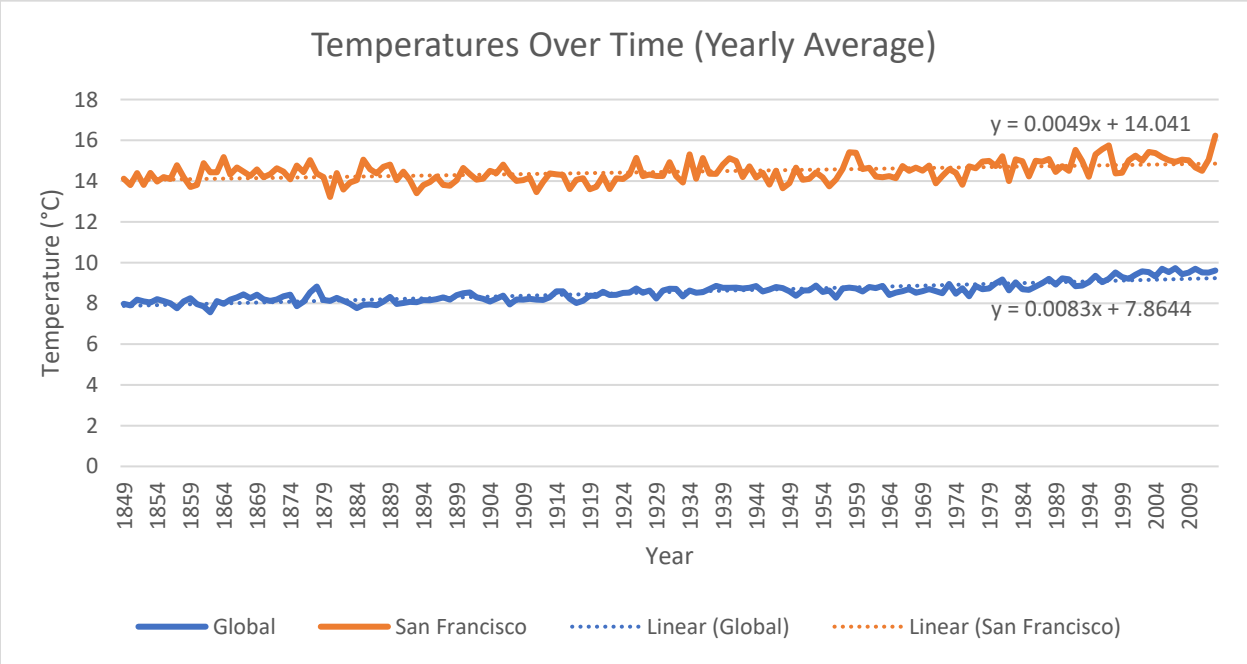


Chart 2

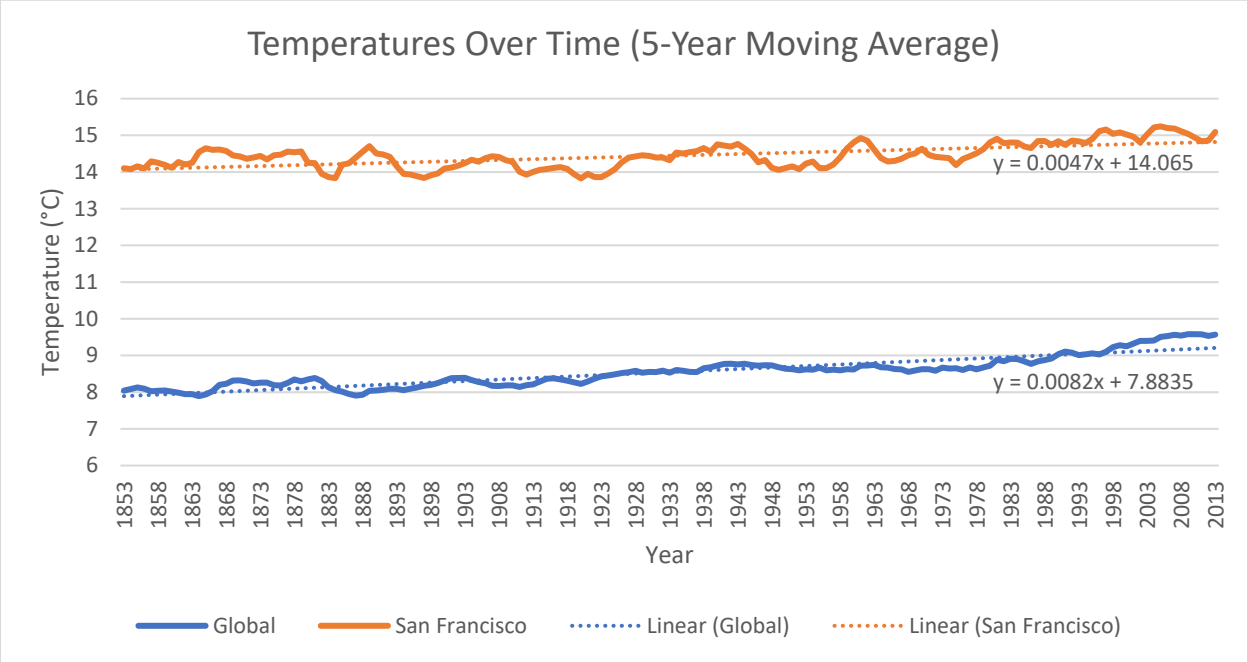


Chart 3

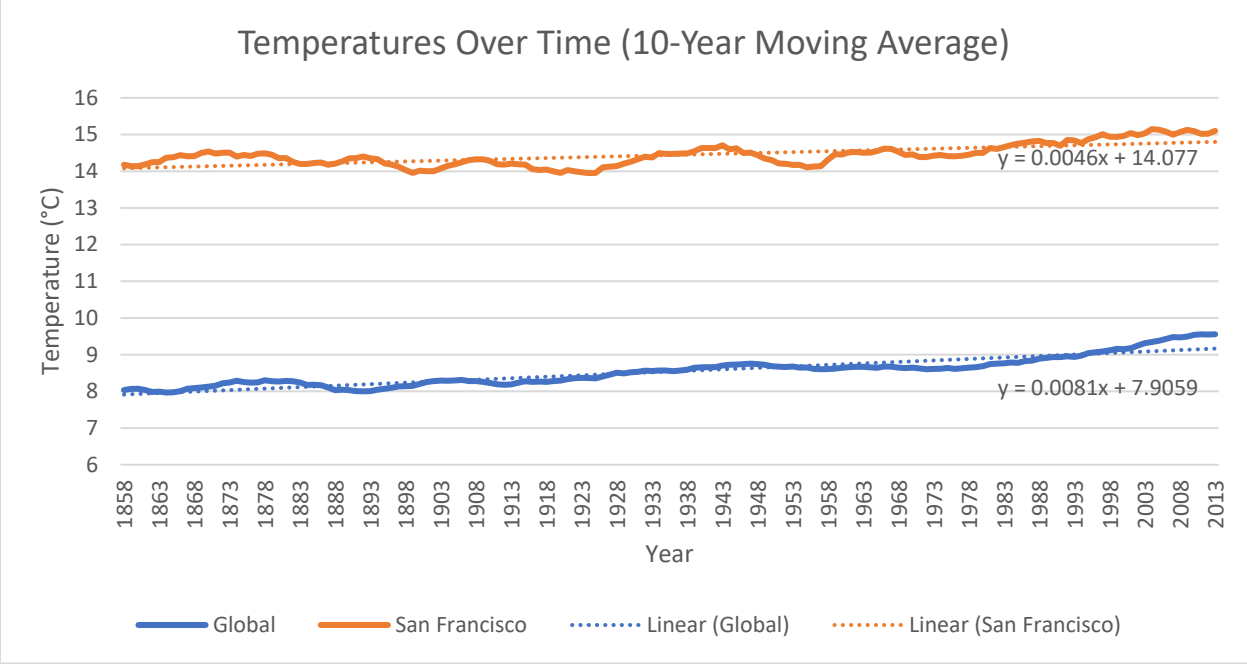


Chart 4

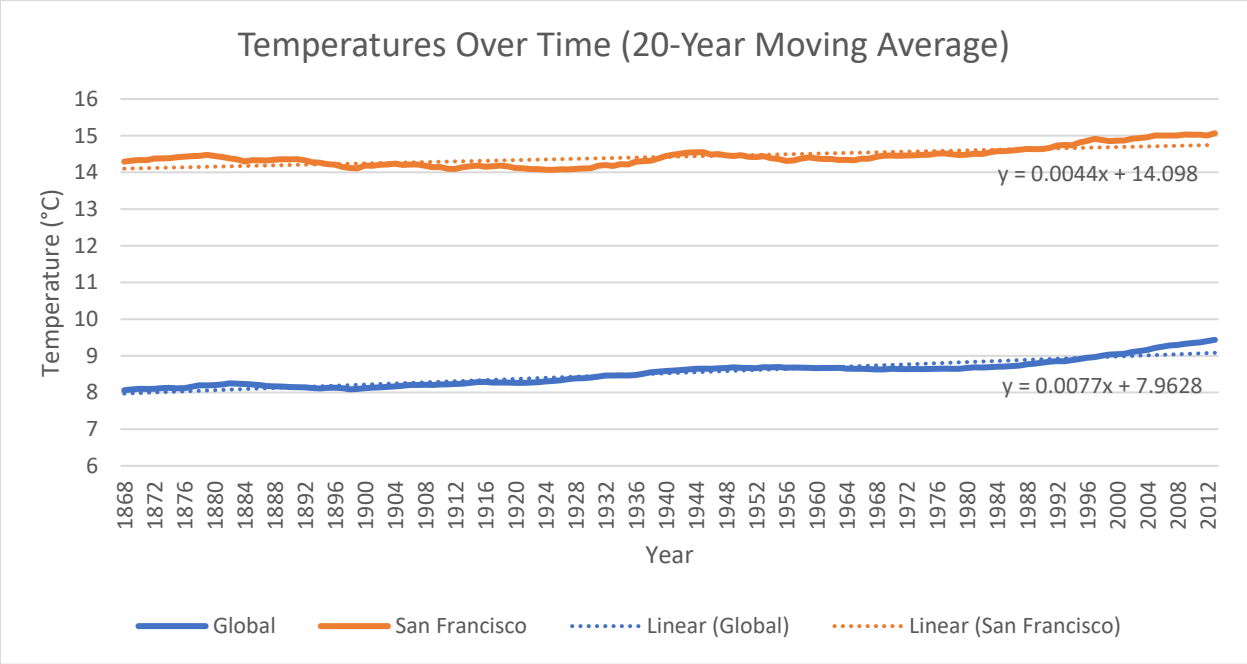
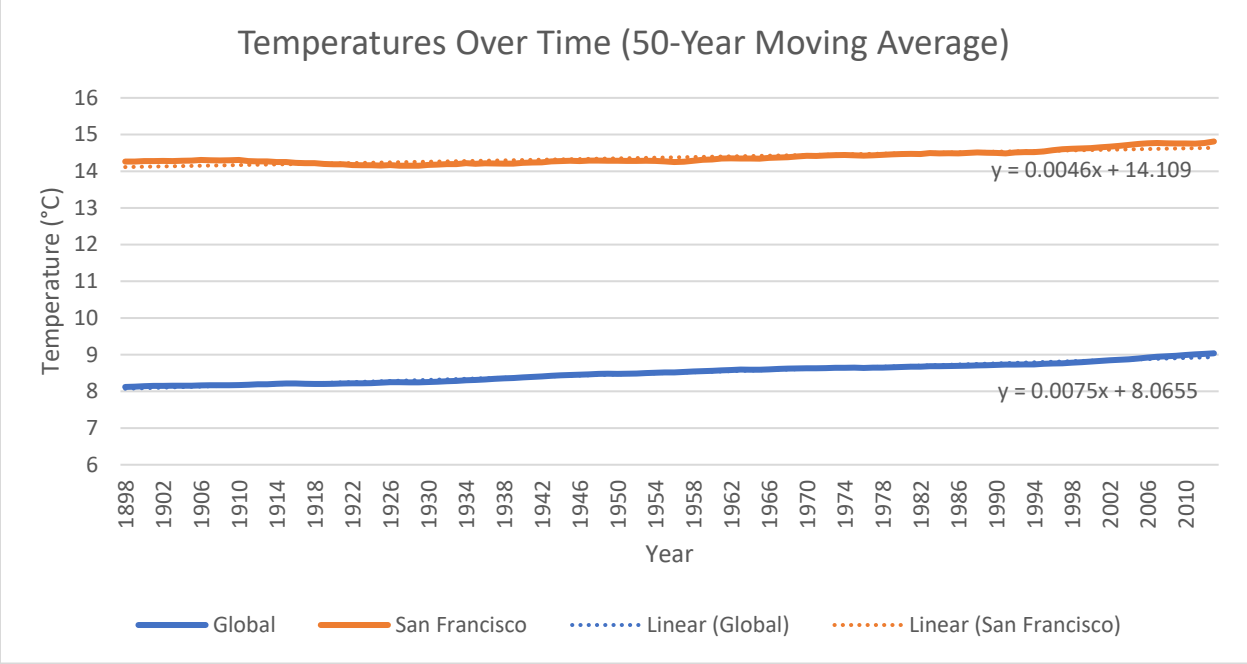


Chart 5



Observations

1. San Francisco has a higher average temperature than the rest of the world.
2. Both Global and San Francisco temperatures have trended upward.
3. San Francisco temperatures have been less consistent than global temperatures.
    a. Caveat: this is could be due to a smaller sample size.
4. The slope of the global trendline decreases as the moving average captures more points. For example, at the yearly average the slope is 0.83%, but at the 50-year moving average the slope is 0.75%.
5. There is not a discernable pattern for the San Francisco trendline as the moving average captures more points.
6. San Francisco's trendline consistently has a smaller slope than Global's. This means that the average temperature in San Francisco has not increased as much as globally over time.