UDACITY

Exploring Weather Trends

Abdullah Yusuf Alhussain

September 28, 2022

1- Data Extraction:

First step is data extraction from the database. In this step, SQL is used. In the beginning, the following query was used to know which Saudi cities are available in the database:

SELECT city, country FROM city_list WHERE country LIKE 'Saudi Arabia';

Results showed that the data of two cities in Saudi Arabia are available in the database, which are Mecca and Riyadh. I am living in eastern province. Hence, the closest city is Riyadh. Now, to extract Riyadh data, the following query is used:

SELECT year, city, country, avg_temp FROM city_data WHERE city LIKE 'Riyadh';

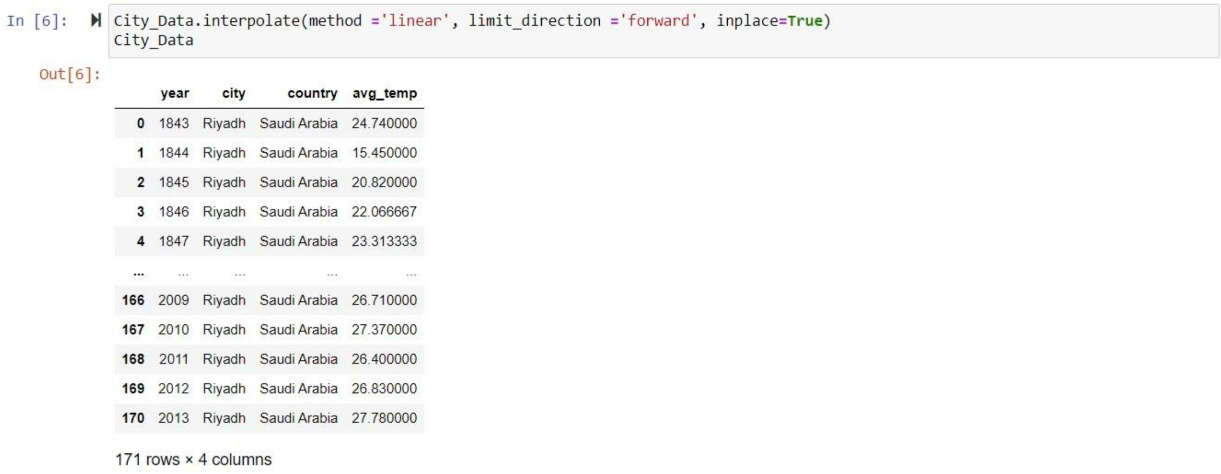
For global data, the following query is used:

SELECT year, avg_temp FROM global_data;

Both datasets were downloaded on my computer.

2- Data Manipulation (Calculating Moving Average):

In this step, Python is used to manipulate the data. There are two issues that need to be resolved before calculating the moving average: First, Riyadh data has two missing values in year 1846 and year 1847. Using linear interpolation, these two missing values were filled as shown in Figure 1.



```
In [6]:  City_Data.interpolate(method ='linear', limit_direction ='forward', inplace=True)
         City_Data
```

```
Out[6]:
        year  city    country       avg_temp
   0    1843  Riyadh  Saudi Arabia  24.740000
   1    1844  Riyadh  Saudi Arabia  15.450000
   2    1845  Riyadh  Saudi Arabia  20.820000
   3    1846  Riyadh  Saudi Arabia  22.066667
   4    1847  Riyadh  Saudi Arabia  23.313333
 ...     ...     ...           ...        ...
 166    2009  Riyadh  Saudi Arabia  26.710000
 167    2010  Riyadh  Saudi Arabia  27.370000
 168    2011  Riyadh  Saudi Arabia  26.400000
 169    2012  Riyadh  Saudi Arabia  26.830000
 170    2013  Riyadh  Saudi Arabia  27.780000

171 rows × 4 columns
```

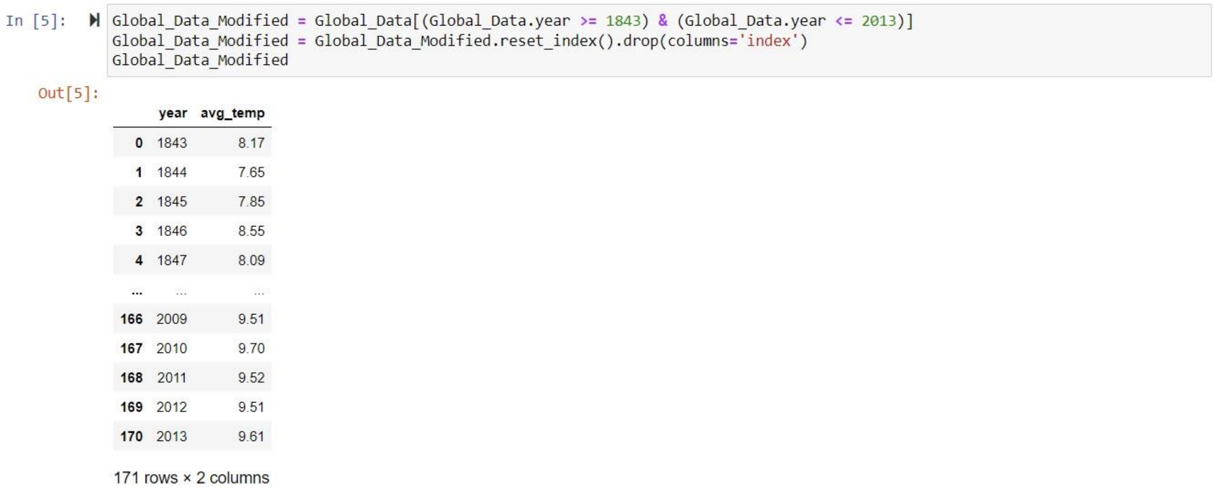
Figure 1: Filling missing values.

Table 1 shows a sample of the data after filling the missing values.

Year	Avg_Temp
1845	20.82
1846	22.067
1847	23.313
1848	24.56

Table 1: Sample data after filling the missing data.

Second, the global data contains 266 data samples where Riyadh data contains only 171 data samples (between year 1843 and 2013). Hence, for the sake of synchronization, the extra samples in the global data are dropped as shown in Figure 2.



In [5]:

```
Global_Data_Modified = Global_Data[(Global_Data.year >= 1843) & (Global_Data.year <= 2013)]
Global_Data_Modified = Global_Data_Modified.reset_index().drop(columns='index')
Global_Data_Modified
```

Out[5]:

	year	avg_temp
0	1843	8.17
1	1844	7.65
2	1845	7.85
3	1846	8.55
4	1847	8.09
...	...	...
166	2009	9.51
167	2010	9.70
168	2011	9.52
169	2012	9.51
170	2013	9.61

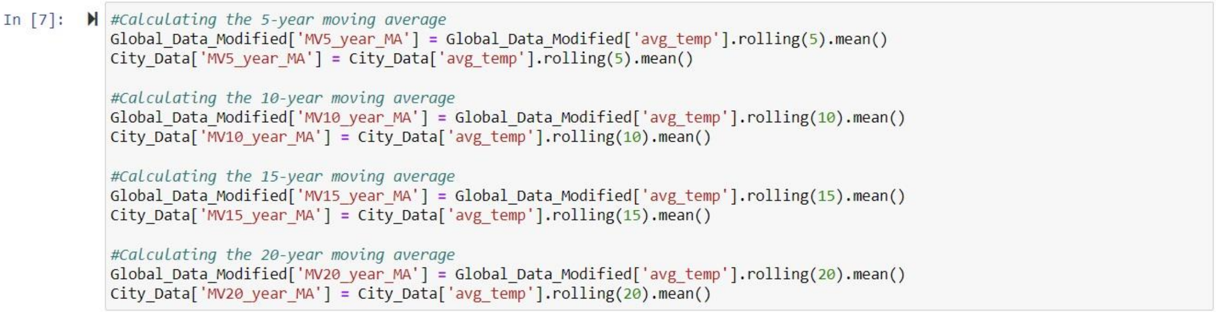
171 rows × 2 columns

Figure 2: Modifying the global data to synchronize it with Riyadh data.

Now, both datasets are clean, synchronized, and ready to be compared with each other. Hence, we can calculate the moving average for both datasets.

The moving average has been calculated four times with a different window size in each time. The window sizes that were considered are 5, 10, 15, and 20. Figure 3 shows the Python code used to implement the moving average calculations for both global data and Riyadh data. Figure 4 shows a plot of global and local temperatures before using moving average (without smoothing). Figures 5 – 8 show the plots of the different implementations of moving average for both datasets. To have a clearer visualization for both plots, it is a good idea to plot them in two separate figures. Also, this will help in visualizing the years that were removed from the global temperature data. Figure 9 – 12 show separate plots of Riyadh temperature and global temperature before and after using the moving average. Clearly, the plots after using the moving average are much smoother and more readable. The best case is the case of the 20-year moving average. Therefore, it is considered for the regression analysis.

A regression line has been fitted to both datasets and the results are shown in the Figures 13 – 14. Since both datasets are nonlinear, fitting a linear regression model is not sufficient to describe the patterns and trends in the data. Hence, a wiser option would be to fit a model that can describe nonlinear data. Piece-wise linear regression has been selected for this purpose. This model fits a line for each interval, so, it approximates each interval with a different line, which will be more accurate than the conventional linear regression model in describing the data. Figures 15 – 16 show both datasets after fitting a piece-wise linear regression model.



```
In [7]:  #Calculating the 5-year moving average
         Global_Data_Modified['MV5_year_MA'] = Global_Data_Modified['avg_temp'].rolling(5).mean()
         City_Data['MV5_year_MA'] = City_Data['avg_temp'].rolling(5).mean()

         #Calculating the 10-year moving average
         Global_Data_Modified['MV10_year_MA'] = Global_Data_Modified['avg_temp'].rolling(10).mean()
         City_Data['MV10_year_MA'] = City_Data['avg_temp'].rolling(10).mean()

         #Calculating the 15-year moving average
         Global_Data_Modified['MV15_year_MA'] = Global_Data_Modified['avg_temp'].rolling(15).mean()
         City_Data['MV15_year_MA'] = City_Data['avg_temp'].rolling(15).mean()

         #Calculating the 20-year moving average
         Global_Data_Modified['MV20_year_MA'] = Global_Data_Modified['avg_temp'].rolling(20).mean()
         City_Data['MV20_year_MA'] = City_Data['avg_temp'].rolling(20).mean()
```

Figure 3: Calculating the moving average.

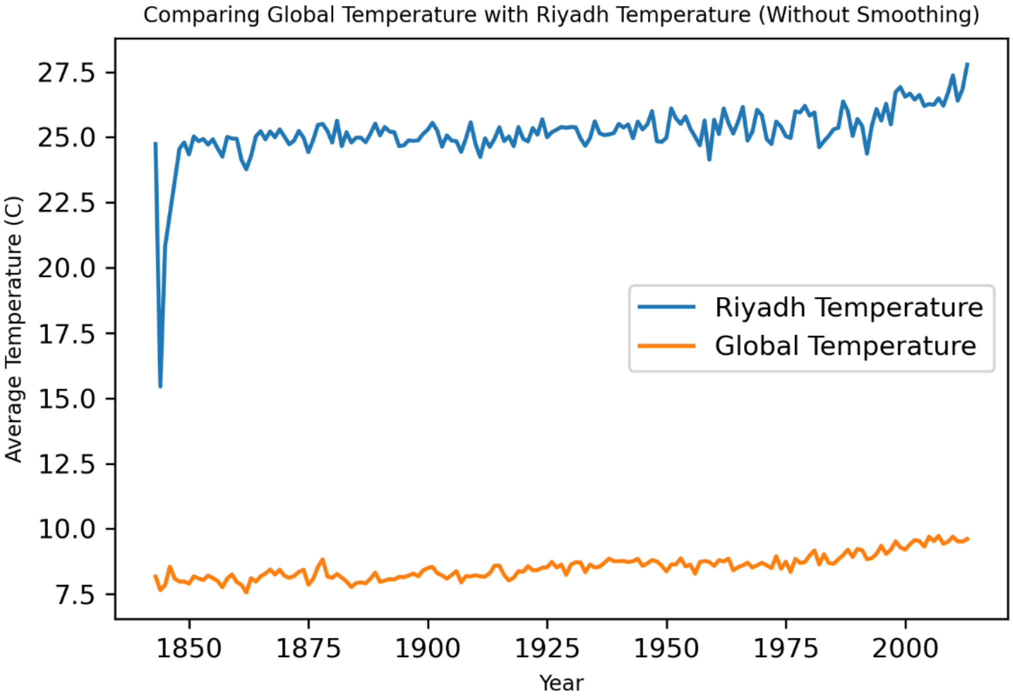


Figure 4: Global temperature and Riyadh temperature without smoothing.

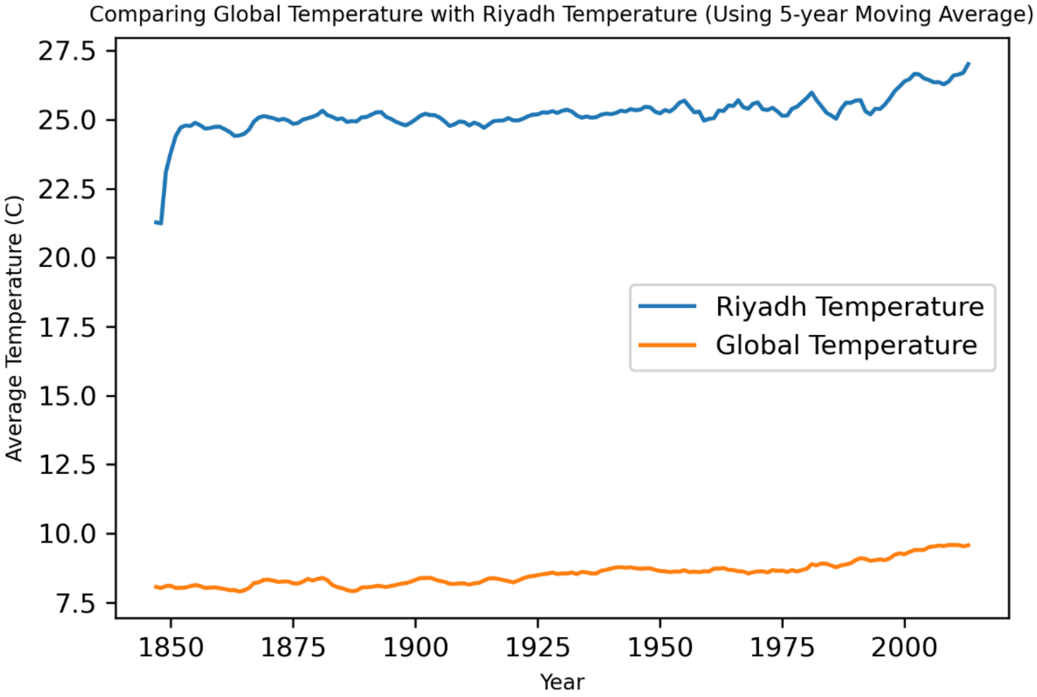


Figure 5: Global temperature and Riyadh temperature after using the 5-year moving average.

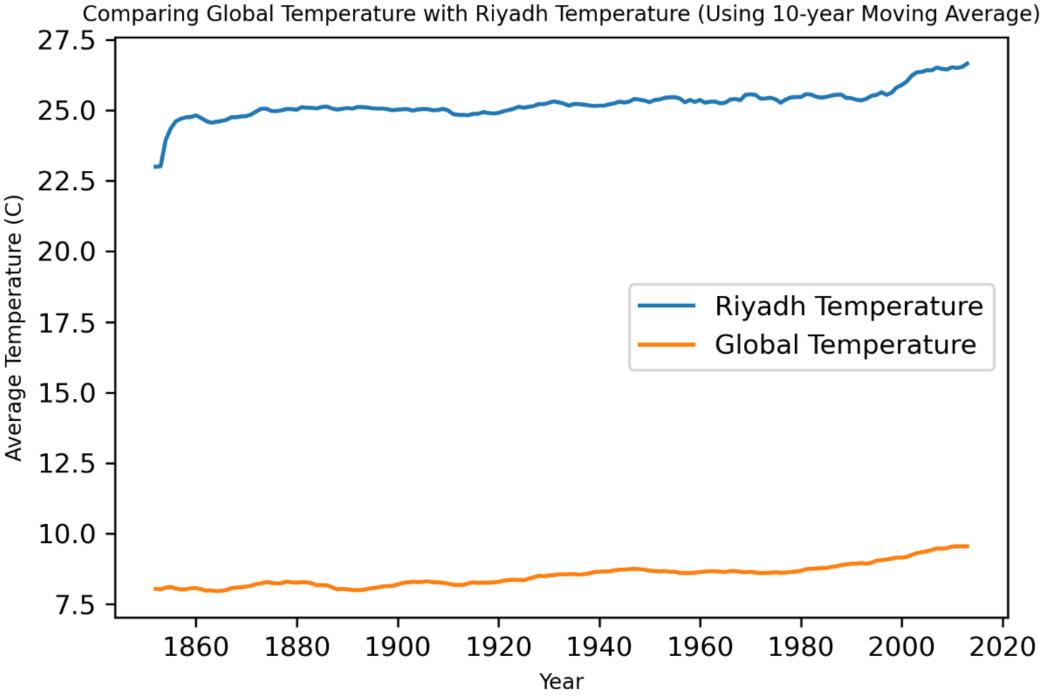


Figure 6: Global temperature and Riyadh temperature after using the 10-year moving average.

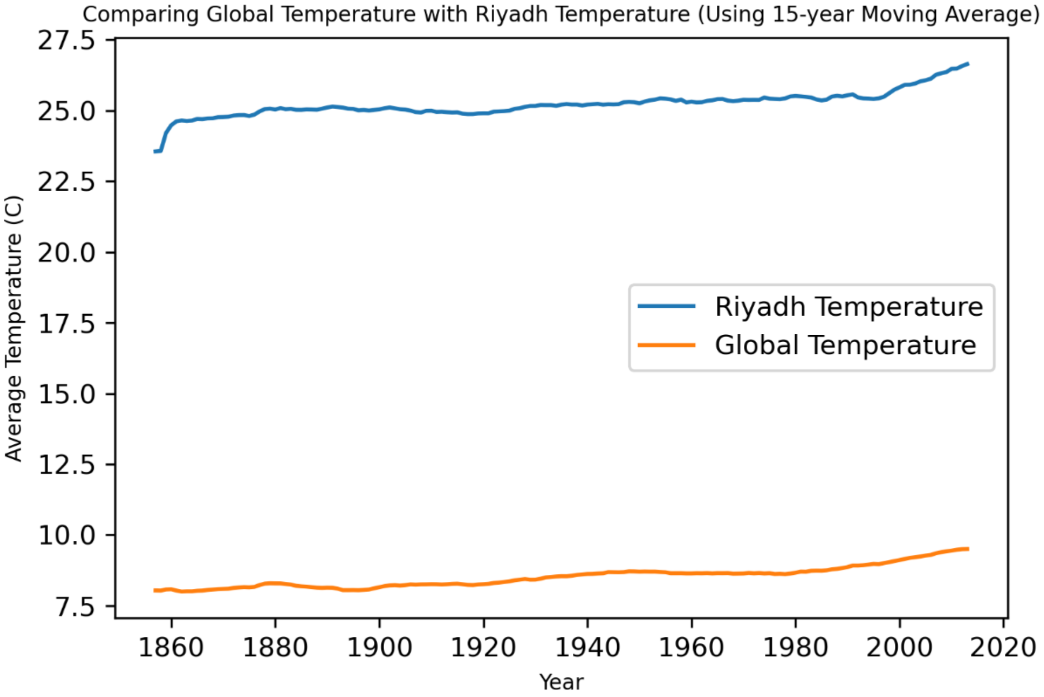


Figure 7: Global temperature and Riyadh temperature after using the 15-year moving average.

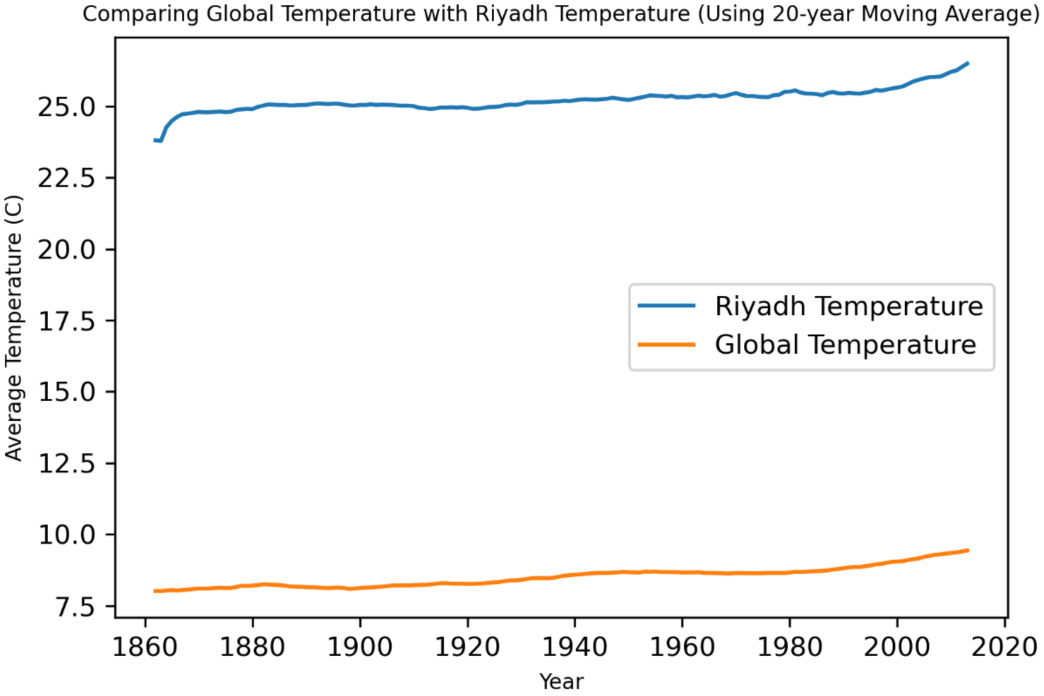


Figure 8: Global temperature and Riyadh temperature after using the 20-year moving average.

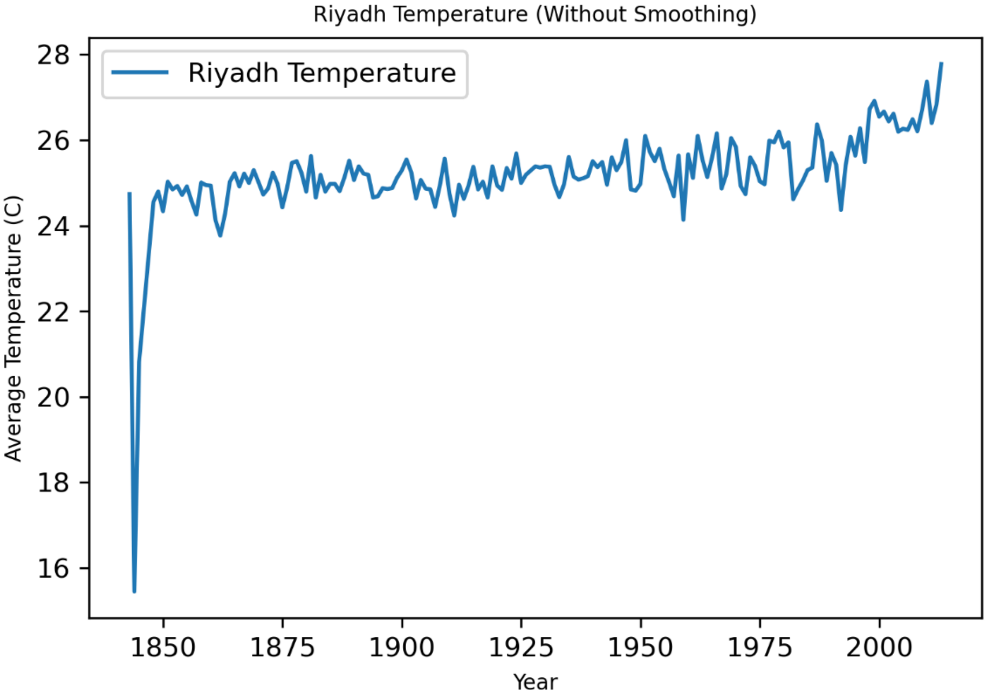


Figure 9: Riyadh temperature (without smoothing).

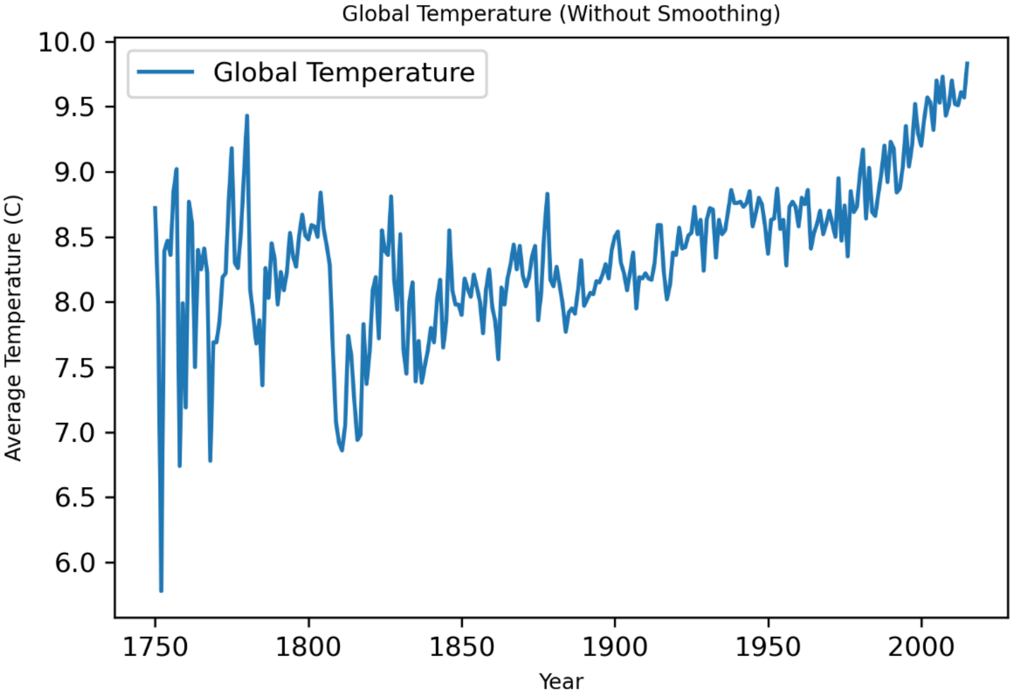


Figure 10: Global temperature (without smoothing).

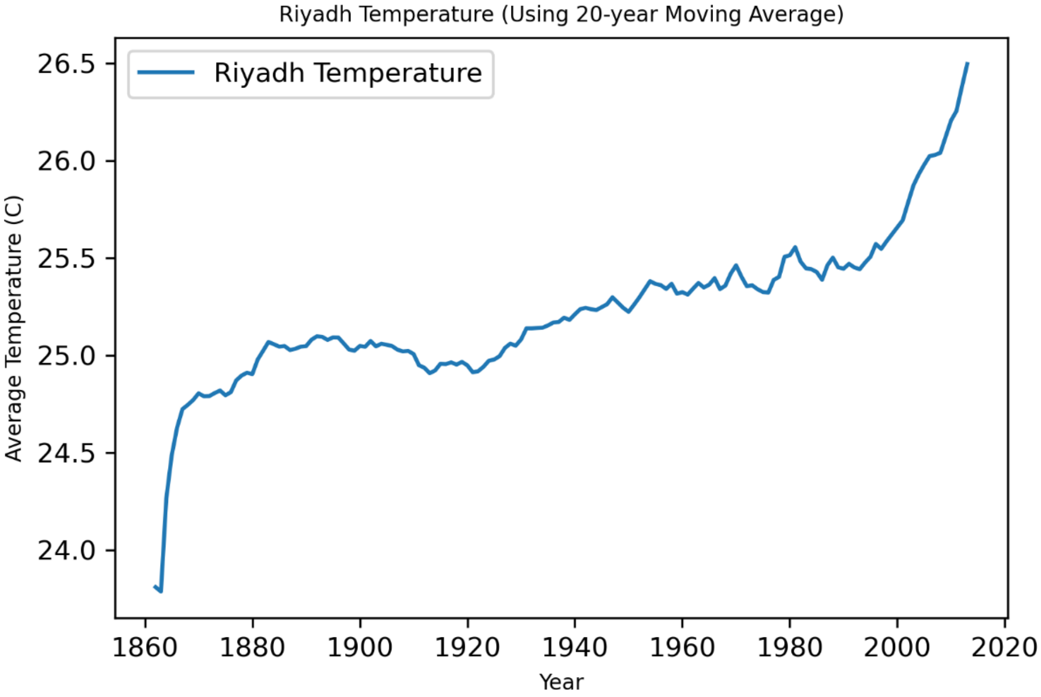


Figure 11: Riyadh temperature after using 20-year moving average.

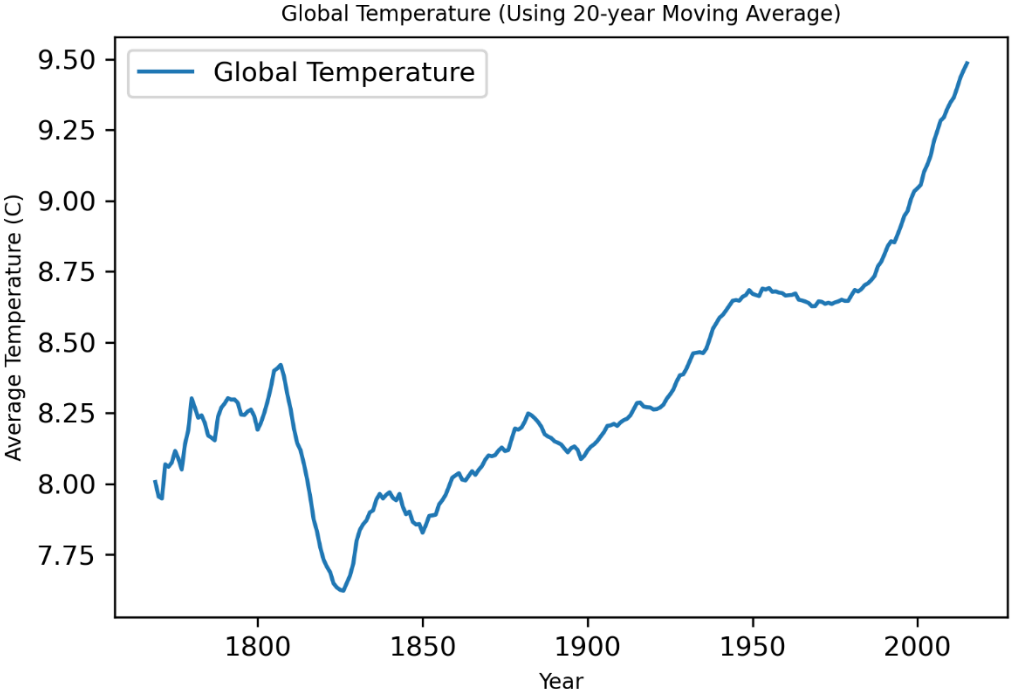


Figure 12: Global temperature after using 20-year moving average.

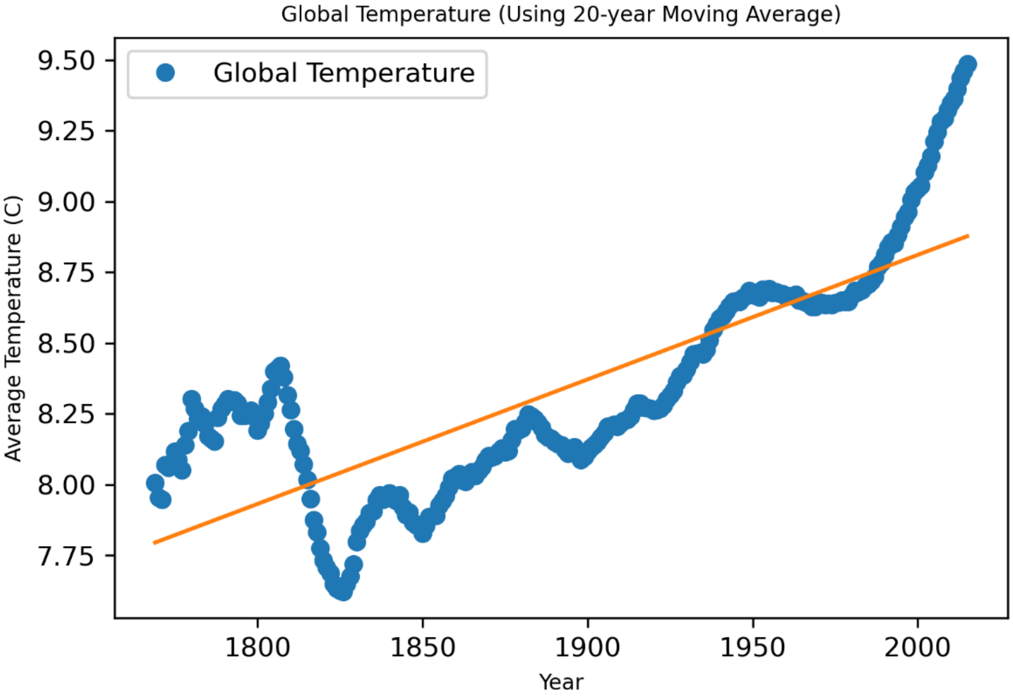


Figure 13: Global temperature after using 20-year moving average with regression line.

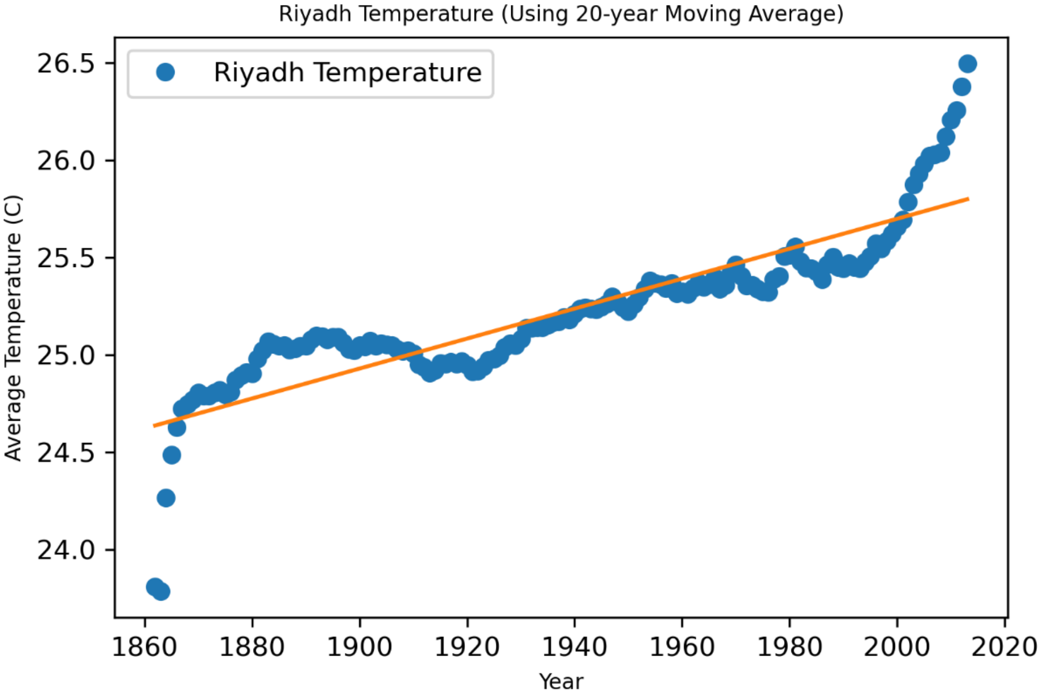


Figure 14: Riyadh temperature after using 20-year moving average with regression line.

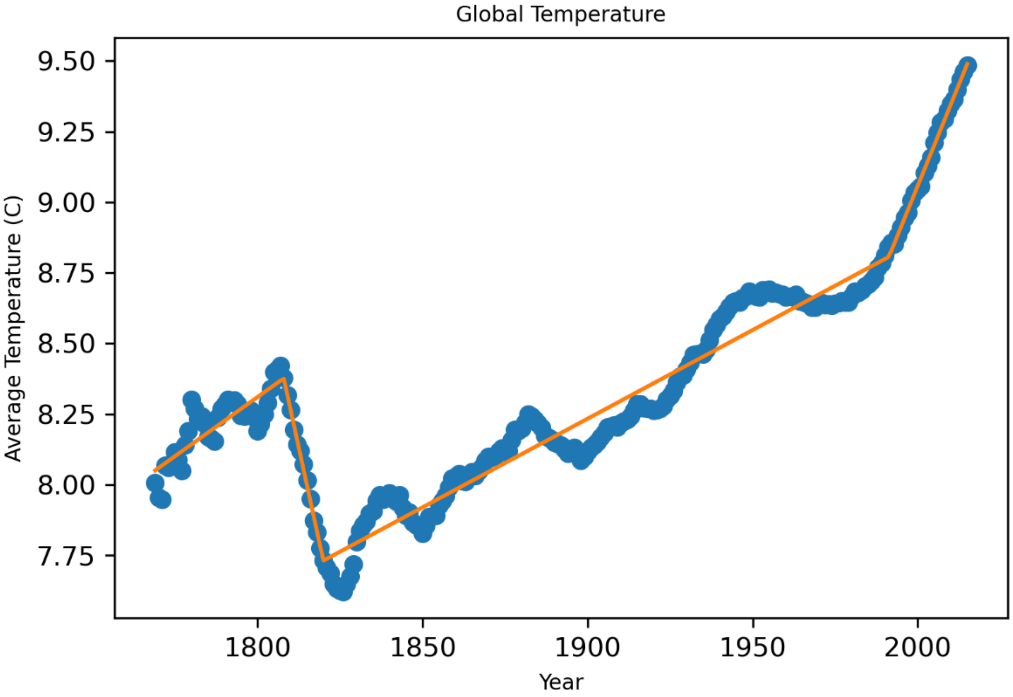


Figure 15: Global temperature after fitting a piece-wise linear regression model.

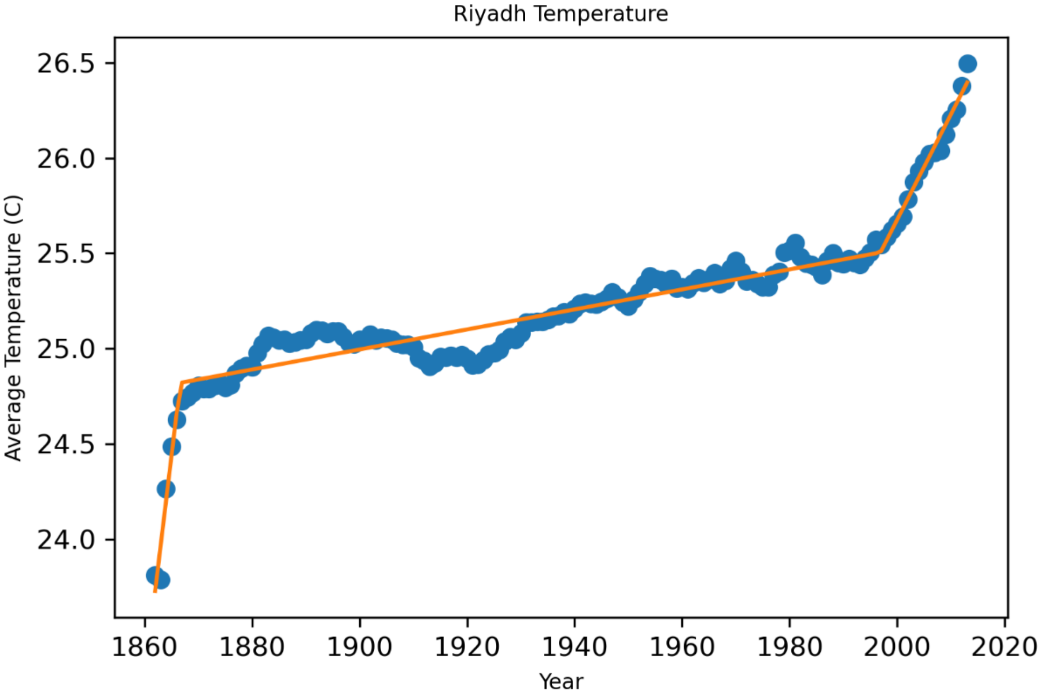


Figure 16: Riyadh temperature after using a piece-wise linear regression model.

Observations:

There are many observations in the data as follows:

1- Riyadh temperature is hotter than the global temperature as can be seen from the Figures in the previous section.

2- The fluctuation in global temperature is higher than the fluctuation in Riyadh temperature even after using the moving average smoothing.

3- The general trends for both local and global temperatures are increasing over time. This can be seen from the lines in Figures 13 – 14. The line equations in the figures are shown in Equation (1) and Equation (2) below. Note: T stands for temperature in degree Celsius (°C) and t is time in years. Also, from the equations (1 and 2), the trend is increasing because both lines have a positive slope.

$$T_{Global}(t) = 0.0044t + 0.00861 \qquad Equation\ (1)$$

$$T_{Riyadh}(t) = 0.0077t + 10.300 \qquad Equation\ (2)$$

4- The temperature data for both datasets is further analyzed by fitting a piece-wise linear regression model as shown in Figures 15 – 16. Both lines can be represented as piece-wise functions shown in equations (3 and 4). This observation will be concentrated on the last 100 years. Before 1990s, the global temperature was increasing at a faster rate than Riyadh temperature since it has a larger slope. However, in recent years, i.e., from around 2000 onward, Riyadh temperature is growing faster because it has a larger slope. Also, for both local and global temperatures, the slope in recent years is much greater than the slope before 1990s. This could be due to the increased number of power plants, factories, cars, and vehicles around the world. The industrial revolution has a negative impact on the climate of the earth.

$$T_{Global}(t) = \begin{cases} 0.00838t - 6.77, & for\ 1769 < t < 1807.95 \\ -0.0539t + 105.81, & for\ 1807.95 < t < 1819.94 \\ 0.00623t - 3.69, & for\ 1819.4 < t < 1991.27 \\ 0.02877t - 48.477, & for\ 1991.27 < t < 2015 \end{cases} \qquad Equation\ (3)$$

$$T_{Riyadh}(t) = \begin{cases} 0.2338t - 411.58, & for\ 1862 < t < 1866.67 \\ 0.005258t + 15, & for\ 1866.67 < t < 1996.87 \\ 0.05534t - 85, & for\ 1996.87 < t < 2013 \end{cases} \qquad Equation\ (4)$$

Questions:

- **What tools did you use for each step? (Python, SQL, Excel, etc)**

As explained in previous sections, SQL was used to extract the data from the database and Python was used to manipulate the data (calculate moving average) and visualize the data.

- **How did you calculate the moving average?**

The code used to calculate the moving average is shown in Figure 3.

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

Three key considerations:

1- Riyadh data had missing values, so, filling these missing values is needed to have better visualization.
2- The time frame of global data is different from Riyadh data time frame. Hence, synchronizing both data frames was needed to visualize both datasets on the same figure.
3- The starting point is adjusted after calculating the moving average as follows:
    - Originally, the data starts from 1843.
    - After calculating the 5-year moving average, it starts from 1847.
    - After calculating the 10-year moving average, it starts from 1852.
    - After calculating the 15-year moving average, it starts from 1857.
    - After calculating the 20-year moving average, it starts from 1862.