# Weather Trends

*RRN*

*8/20/2018*

This is the markdown document of a climate analysis prjoect that I have written in R. The project analyzes the temperature trends over the past few decades in Seattle, WA, compared to the rest of the world. For this project, I first extracted the global temperature data and the temperature data for the city of Seattle as csv files, from the internet. For the data plotting, I have used the ggplot2 package in RStudio version 1.1.456, on the Ubuntu (LINUX) platform. I have included the SQL queries and R-code along with the plotting data in the document.

After extracting the data as data frames and cleaning up missing data, I visualized the raw trends for the city (Seattle) and global data using scatter plots (Figures 1 and 2). The two trends were compared over time. I plotted the scatters together (Figure 3), and observed that the average global temperature recordings run higher than those recorded for Seattle. I confirmed this finding by plotting the total average along with standard deviation, over the years as a barplot for both the variables, Seattle and Global (Figure 4).

To clearly observe the trends, I plotted the Seattle and Global data using line plots (Figure 5). To smoothen the trendlines and avoid ambiguity, I used the rollapply function to calculate and plot the rolling avergaes for the two variables, using a window size of 10 years. The trend shows that while Seattle tends to have lower temperatures than the global averages, the overall temperature trend displays an increase over the years for Seattle and the rest of the World, reflecting the positive trend towards global warming over the 4 centuries. My observations/conclusions are listed towards the bottom of the document.

The SQL queries used to extract the data from the database were as follows:

SELECT * FROM city_list ORDER BY country;

SELECT * FROM global_data;

SELECT * FROM city_data WHERE city = 'Seattle';

After extraction following each query, the data were exported as csv files.

The R-codes along with the plots are shown below.

```r
library (ggplot2)
library (zoo)
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
# Import the Seattle temperature data into a data frame
seattle_data1 = as.data.frame(read.csv("seattle_data.csv"))

any(is.na (seattle_data1))
```

```
## [1] TRUE
```

```r
seattle_data <- seattle_data1[complete.cases(seattle_data1), ]

tail (seattle_data)
```

```
##     year    city       country avg_temp
## 181 2008 Seattle United States     7.57
## 182 2009 Seattle United States     8.02
## 183 2010 Seattle United States     8.25
## 184 2011 Seattle United States     7.35
## 185 2012 Seattle United States     8.08
## 186 2013 Seattle United States     9.95
```

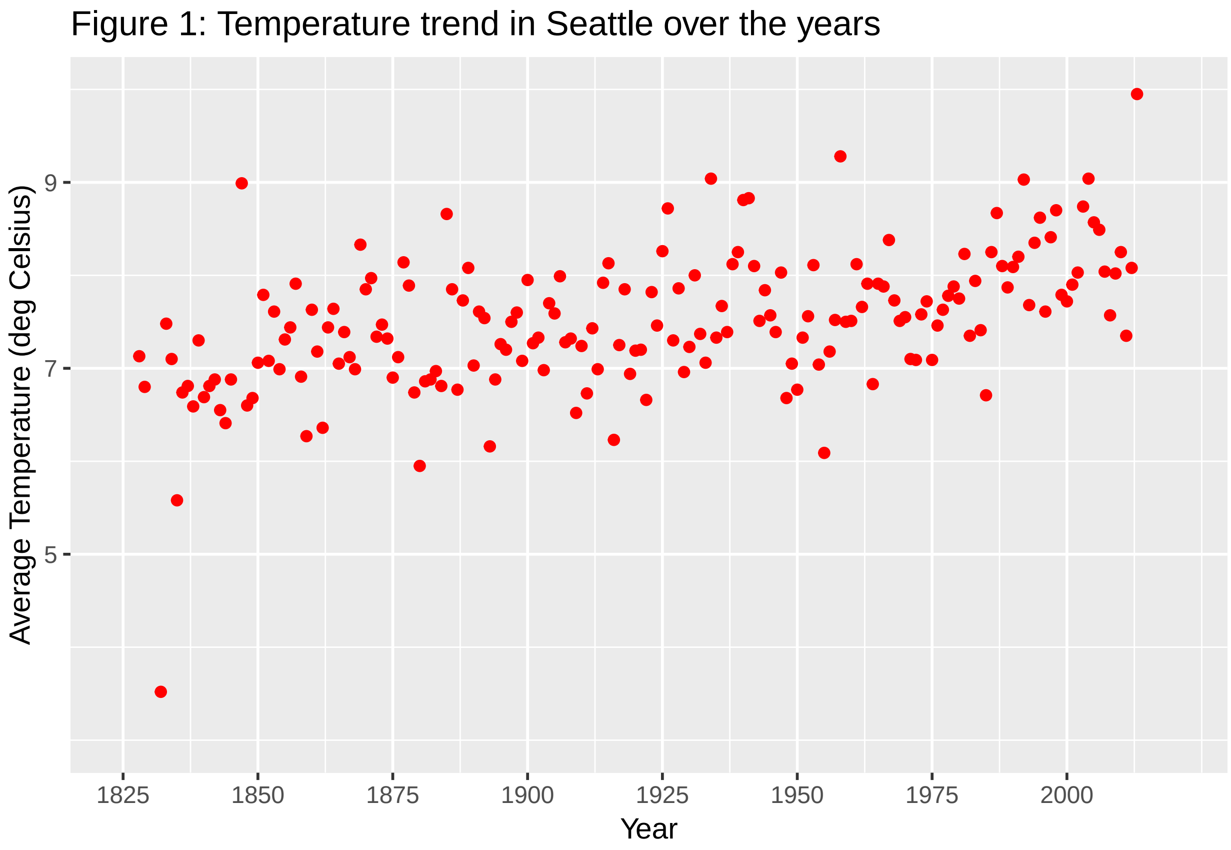
```
str (seattle_data)
```

```
## 'data.frame':    183 obs. of  4 variables:
##  $ year    : int  1828 1829 1832 1833 1834 1835 1836 1837 1838 1839 ...
##  $ city    : Factor w/ 1 level "Seattle": 1 1 1 1 1 1 1 1 1 1 ...
##  $ country : Factor w/ 1 level "United States": 1 1 1 1 1 1 1 1 1 1 ...
##  $ avg_temp: num  7.13 6.8 3.52 7.48 7.1 5.58 6.74 6.81 6.59 7.3 ...
```

```
any(is.na (seattle_data))
```

```
## [1] FALSE
```

```
seattle = ggplot (seattle_data, aes (x=year, y=avg_temp)) +
  geom_point(position = position_dodge(width=1), color = "red") +
  labs (y= "Average Temperature (deg Celsius)", x = "Year", title = "Figure 1: Temperature trend in Sea
  scale_x_continuous(limits = c(1825, 2020), breaks = seq(1825, 2015, 25))+
  scale_y_continuous(limits = c(3, 10), breaks = seq(5, 10, 2))
seattle
```



```
# Import the global temperature data into a data frame
global_data = as.data.frame(read.csv("global_data.csv"))
```

```
any(is.na (global_data))
```

```
## [1] FALSE
```

```
head (global_data)
```

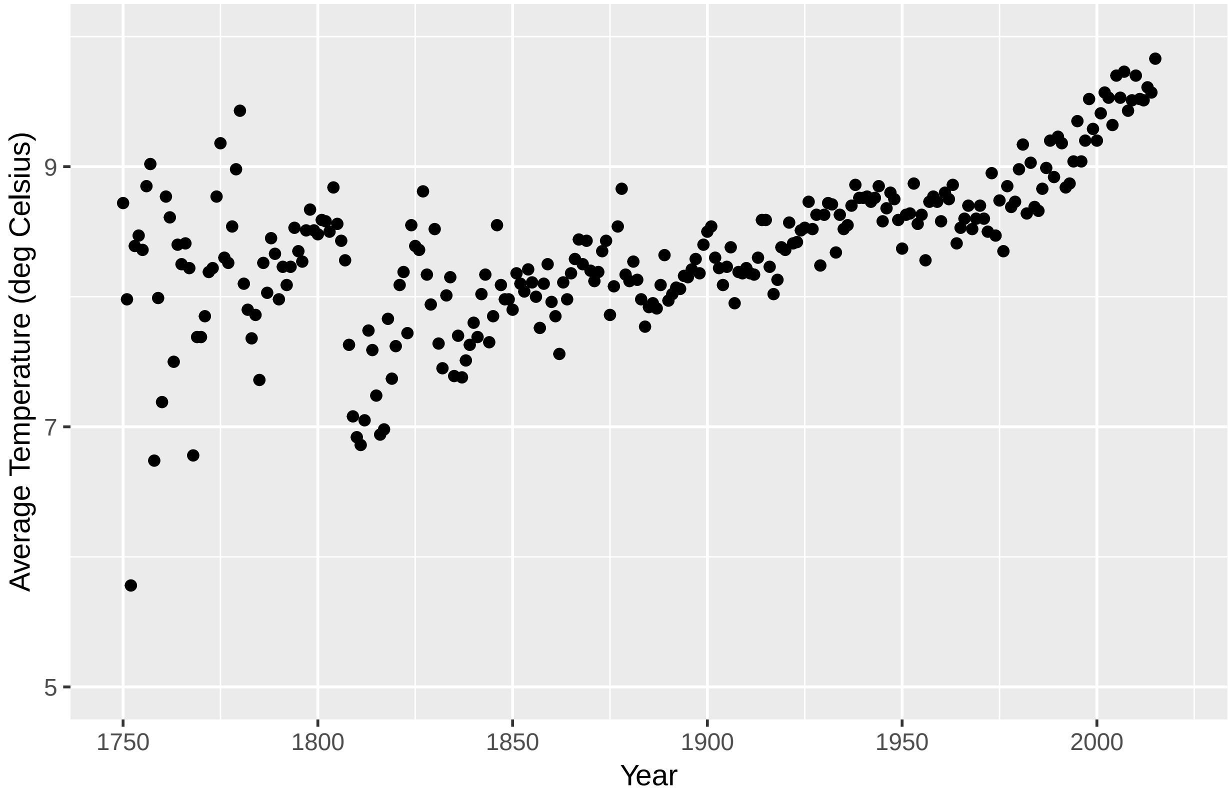
```
##   year avg_temp
## 1 1750     8.72
## 2 1751     7.98
## 3 1752     5.78
## 4 1753     8.39
## 5 1754     8.47
## 6 1755     8.36
```

```
str (global_data)
```

```
## 'data.frame':    266 obs. of  2 variables:
##  $ year    : int  1750 1751 1752 1753 1754 1755 1756 1757 1758 1759 ...
##  $ avg_temp: num  8.72 7.98 5.78 8.39 8.47 8.36 8.85 9.02 6.74 7.99 ...
```

```
# Plot the global temperature data
global = ggplot (global_data, aes (x=year, y=avg_temp)) +
  geom_point(position = position_dodge(width=1)) +
  labs (y= "Average Temperature (deg Celsius)", x = "Year", title = "Figure 2: Temperature trend across
  scale_x_continuous(limits = c(1750, 2020), breaks = seq(1750, 2015, 50))+
  scale_y_continuous(limits = c(5, 10), breaks = seq(5, 10, 2))
global
```

Figure 2: Temperature trend across the globe over the years



```
# Import the global and Seattle temperature data into a data frame
global_seattle_data1 = as.data.frame(read.csv("global_seattle_data.csv"))
any(is.na(global_seattle_data1))
```

```
## [1] TRUE
```

```
global_seattle_data = na.omit(global_seattle_data1)

head (global_seattle_data)
```

```
##   Year Region AvgTemp
## 1 1750 Global    8.72
## 2 1751 Global    7.98
## 3 1752 Global    5.78
## 4 1753 Global    8.39
## 5 1754 Global    8.47
## 6 1755 Global    8.36
```

```
str (global_seattle_data)
```

```
## 'data.frame':    449 obs. of  3 variables:
##  $ Year   : int  1750 1751 1752 1753 1754 1755 1756 1757 1758 1759 ...
##  $ Region : Factor w/ 2 levels "Global","Seattle": 1 1 1 1 1 1 1 1 1 1 ...
##  $ AvgTemp: num  8.72 7.98 5.78 8.39 8.47 8.36 8.85 9.02 6.74 7.99 ...
##  - attr(*, "na.action")= 'omit' Named int  267 268 269 270 271 272 273 274 275 276 ...
##   ..- attr(*, "names")= chr  "267" "268" "269" "270" ...
```

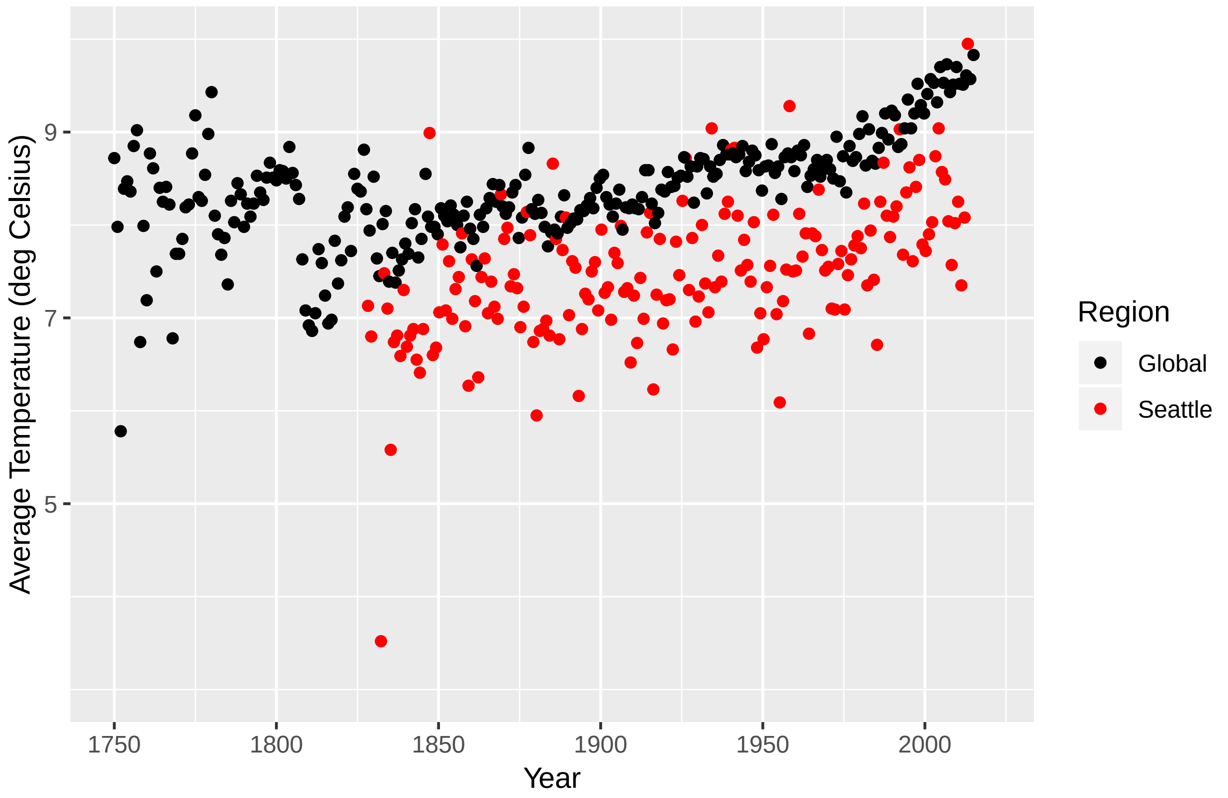
```
any(is.na(global_seattle_data))
```

```
## [1] FALSE
```

```
# Plot the global and Seattle scatters together for comparisons

global_seattle = ggplot (global_seattle_data, aes(x=Year, y=AvgTemp, color = Region)) +
  geom_point (position = position_dodge(width=1)) +
  labs (y= "Average Temperature (deg Celsius)", x = "Year", title = "Figure 3: Temperature trend in Sea
  scale_color_manual(values = c("black", "red")) +
  scale_x_continuous(limits = c(1750, 2020), breaks = seq(1750, 2015, 50))+
  scale_y_continuous(limits = c(3, 10), breaks = seq(5, 10, 2))
global_seattle
```

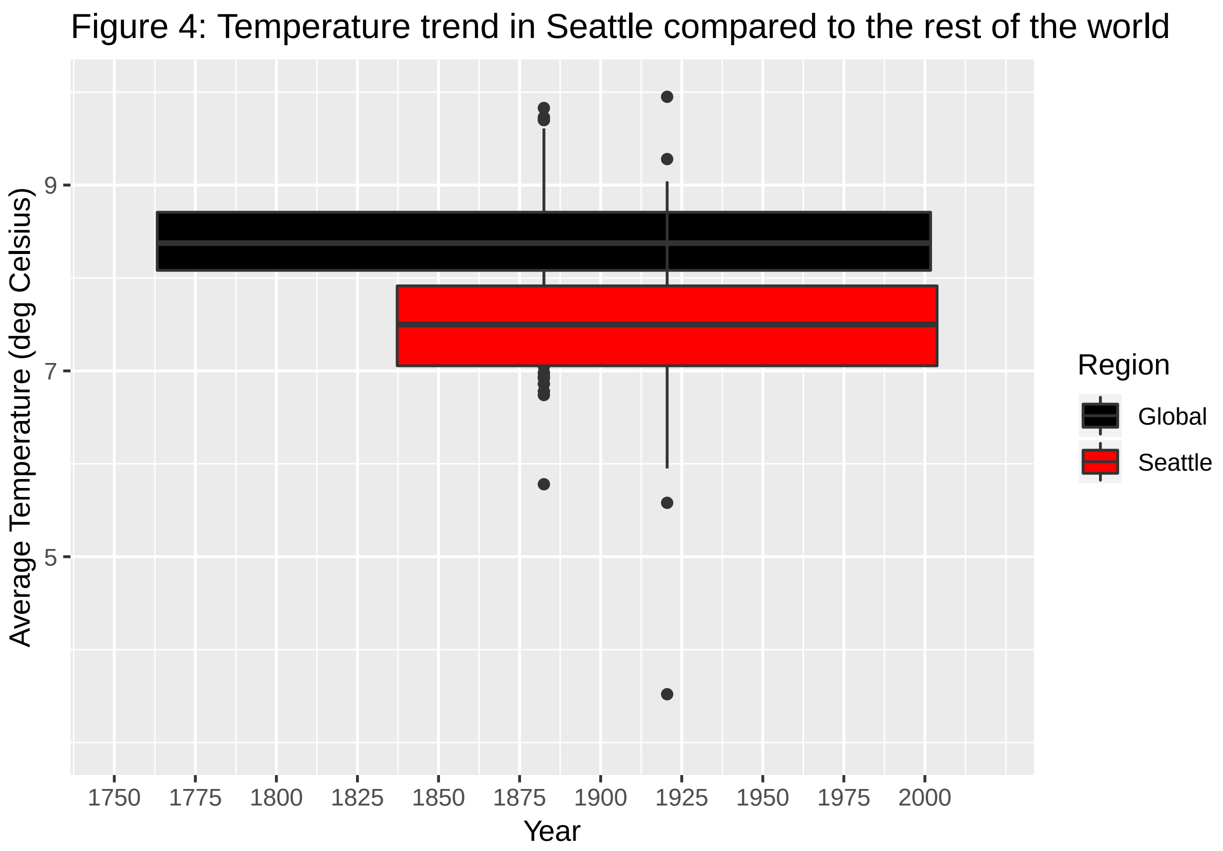
Figure 3: Temperature trend in Seattle compared to the rest of the world



```
# Plot the global and Seattle scatters together for comparisons (see outliers better)

global_seattle_box = ggplot (global_seattle_data, aes(x=Year, y=AvgTemp, fill = Region))+
  geom_boxplot (position = position_dodge(width=0.85)) +
  labs (y= "Average Temperature (deg Celsius)", x = "Year", title = "Figure 4: Temperature trend in Sea
  scale_fill_manual(values = c("black", "red")) +
  scale_x_continuous(limits = c(1750, 2020), breaks = seq(1750, 2015, 25))+
  scale_y_continuous(limits = c(3, 10), breaks = seq(5, 10, 2))
global_seattle_box
```

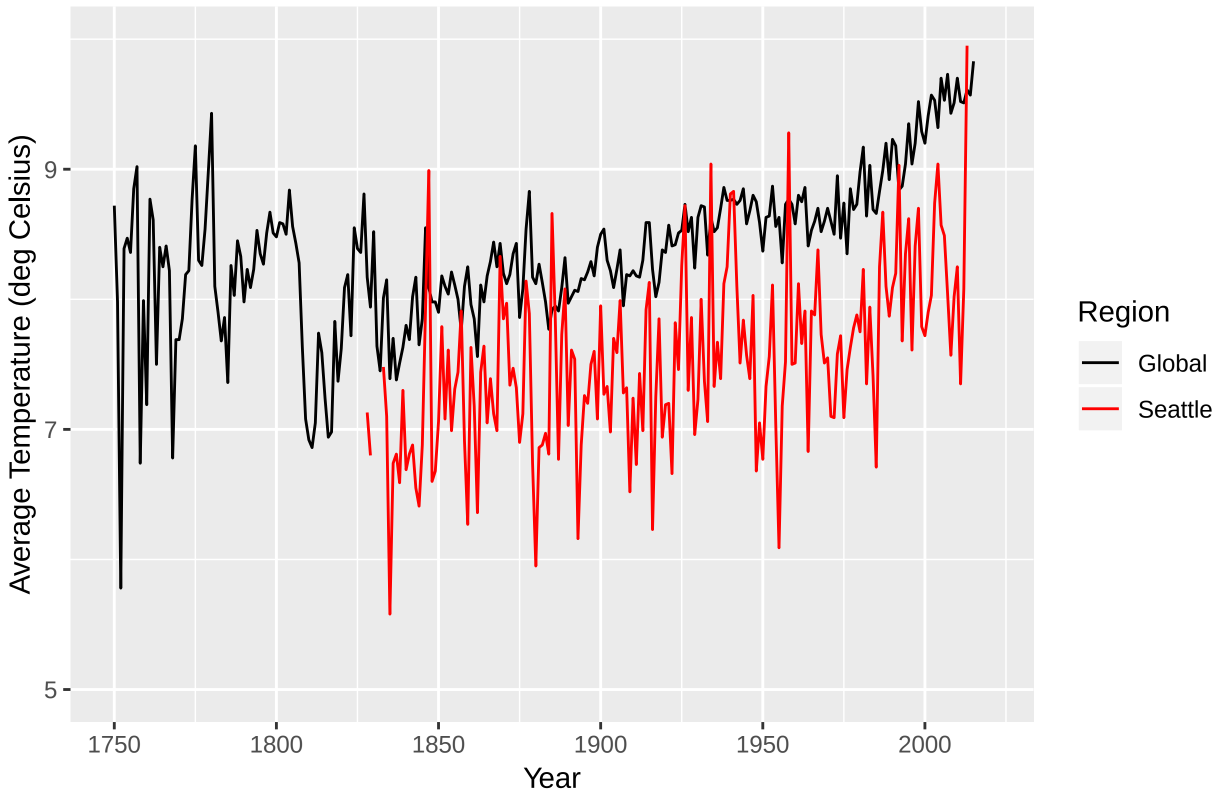
```
## Warning: position_dodge requires non-overlapping x intervals
```



```
# This gives a warning. Ideally, for the boxplot, I should plot the categorical variables on X-axis. I
```

```
# Plot the global and Seattle data together as line plots
global_seattle = ggplot (global_seattle_data, aes(x=Year, y=AvgTemp, color = Region))+
  geom_line () +
  labs (y= "Average Temperature (deg Celsius)", x = "Year", title = "Figure 5: Temperature trend in Sea
  scale_color_manual(values = c("black", "red")) +
  scale_x_continuous(limits = c(1750, 2020), breaks = seq(1750, 2015, 50))+
  scale_y_continuous(limits = c(5, 10), breaks = seq(5, 10, 2))
global_seattle
```

Figure 5: Temperature trend in Seattle compared to the rest of the world



```
# Plot the global and Seattle data together as rolling averages to smoothen the plot
global_seattle_sdata1 = as.data.frame(read.csv("global_seattle_data_side.csv"))
any(is.na(global_seattle_sdata1))
```

```
## [1] TRUE
```

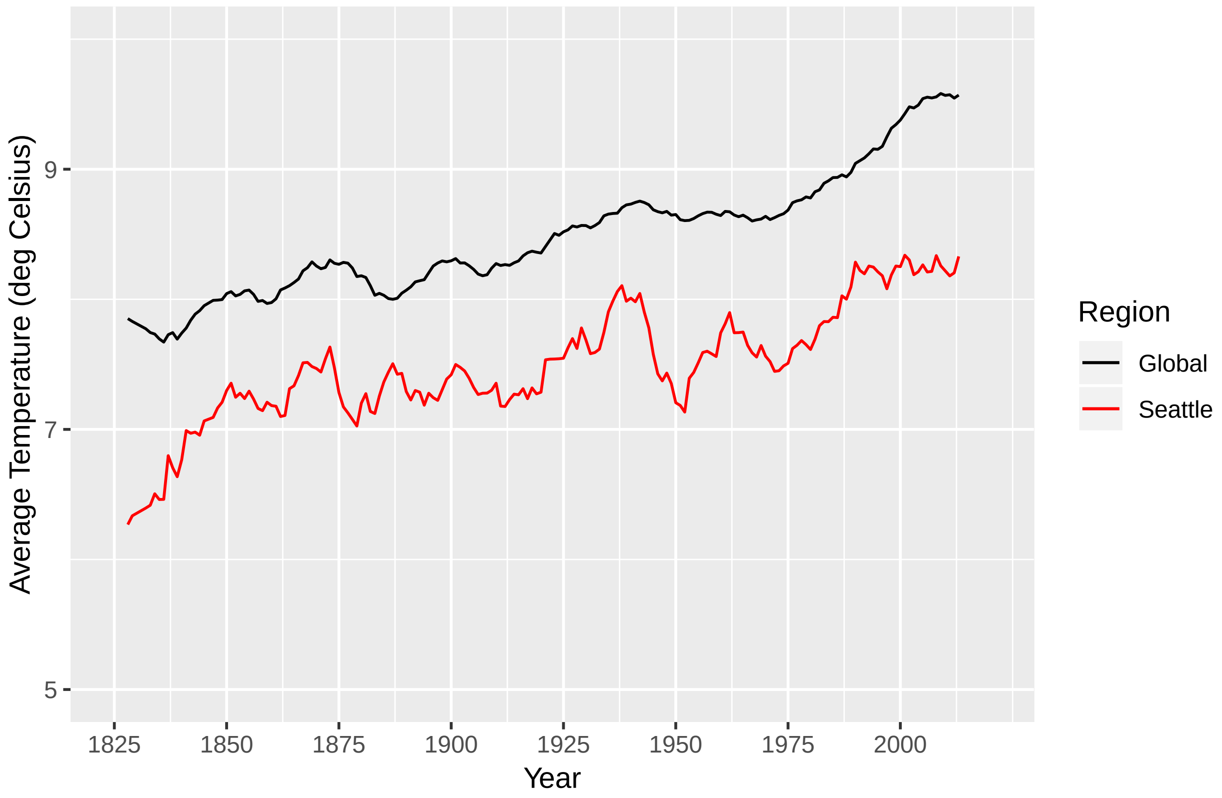
```
global_seattle_sdata = na.omit(global_seattle_sdata1)
any(is.na(global_seattle_sdata))
```

```
## [1] FALSE
```

```
global_seattle_sdata$Global = rollapply (global_seattle_sdata$Global, 10, mean, fill = list (NA, NA, NA)
global_seattle_sdata$Seattle = rollapply (global_seattle_sdata$Seattle, 10, mean, fill = list (NA, NA, N
global_seattle1 = ggplot (global_seattle_sdata, aes(Year))+
  geom_line (aes(y=Global, color= "Global")) +
  geom_line (aes(y=Seattle, color= "Seattle")) +
  labs (y= "Average Temperature (deg Celsius)", x = "Year", title = "Figure 6: Temperature trend in Sea
  scale_color_manual(values = c("black", "red")) +
  scale_x_continuous(limits = c(1825, 2020), breaks = seq(1825, 2015, 25))+
  scale_y_continuous(limits = c(5, 10), breaks = seq(5, 10, 2))

global_seattle1
```

Figure 6: Temperature trend in Seattle compared to the rest of the world



My observations about the data:

1. The data for the global average temperature is available for a longer period than for the Seattle data (1750 to 2015 for Global versus 1828 to 2013 for Seattle).
2. The total average and overall trend over the years reflect the fact that the global temperature trend runs higher than that for the city of Seattle.
3. The highest recorded temperature for the globe was 9.83 in the year 2015, while the lowest was 5.78 in the year 1752. The highest temperature recorded for Seattle was 9.95 in the year 2013, while the lowest was 3.52 in the year 1832.
4. A comparison of the trends shows that the average temperatures for Seattle and the World has been increasing over the years. Despite the rare outliers, the data prove the positive trend towards global warming over the 4 centuries.