Exploring Weather Trends

Introduction

The goal of this project was to analyze and compare global temperature data and the local temperature data of Victoria, BC, Canada. Visualizations of both temperature data were created to compare the overall trends between world temperture and the temperature of Victoria.

Data Source

The data were stored on the SQL server at Udacity. The database consisted of three tables: city_data, city_list, and global_data.

First, the global temperature data were extracted from the table global_data by using the below SQL query:

```sql
SELECT *
FROM global_data;
```

The extracted data were then exported to the file `world.csv`.

Next, to identify whether the city_data contained local temperature data of Vancouver, or any nearby major cities (i.e. Victoria, Seattle), another SQL query was written to search for matching city names in city_list:

```sql
SELECT *
FROM city_list
WHERE city IN ('Vancouver','Victoria','Seattle');
```

The results showed that only Victoria and Seattle were in city_list. The local temperature data of Victoria were then extracted from city_data using the below SQL query and then exported to the file `global.csv`:

```sql
SELECT *
FROM city_data
WHERE city = 'Victoria';
```

Data Exploration

After the csv files were properly extracted from the SQL server, R was used to explore, analyze, and visualize the temperature data.

```r
world_temp <- read.csv('world.csv')
victoria_temp <- read.csv('victoria.csv')
```

To get a better sense of what the data looked like, the first 10 rows of the world data were shown below:

```r
head(world_temp, 10)
```

```
##    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
## 7  1756     8.85
## 8  1757     9.02
## 9  1758     6.74
## 10 1759     7.99
```

and the first 10 rows of victoria data:

```r
head(victoria_temp, 10)
```

```
##    year     city country avg_temp
## 1  1828 Victoria  Canada     6.83
## 2  1829 Victoria  Canada     6.58
## 3  1830 Victoria  Canada       NA
## 4  1831 Victoria  Canada       NA
## 5  1832 Victoria  Canada     3.25
## 6  1833 Victoria  Canada     7.27
## 7  1834 Victoria  Canada     6.81
## 8  1835 Victoria  Canada     5.35
## 9  1836 Victoria  Canada     6.52
## 10 1837 Victoria  Canada     6.61
```

There are missing values in the Victoria data, and could be problematic when calculating the moving avareges. The proportion of missing values in this data is:

```r
sum(is.na(victoria_temp$avg_temp)) / nrow(victoria_temp)
```

```
## [1] 0.01612903
```

Roughly about 1.6%, which was acceptable. The strategy that was employed in this exploration was to replace the missing values by linear interpolation. The library `imputeTS` has a function, `na.interpolate()` that provides an easy way to do linear interpolation.

```r
library(imputeTS)
victoria_temp$avg_temp <- as.numeric(na.interpolation(victoria_temp$avg_temp, 'linear'))
```

Looking at the first few rows of Victoria data again:

```r
head(victoria_temp, 6)
```

```
##   year     city country avg_temp
## 1 1828 Victoria  Canada     6.83
## 2 1829 Victoria  Canada     6.58
## 3 1830 Victoria  Canada     5.47
## 4 1831 Victoria  Canada     4.36
## 5 1832 Victoria  Canada     3.25
## 6 1833 Victoria  Canada     7.27
```

The missing values were now properly replaced.

Moving Averages

10-, 20-, and 50-year moving averages were calculated to show the short, mid, and long term changes in the global and Victoria temperatures. The `filter()` function in base R provides an easy to calculate moving averages. To calculate the moving averages for global data:

```r
#10-day moving average
world_temp$mav10 <- as.numeric(stats::filter(world_temp$avg_temp,
                                             rep(1/10, 10), side = 1))
#20-day moving average
world_temp$mav20 <- as.numeric(stats::filter(world_temp$avg_temp,
                                             rep(1/20, 20), side = 1))
#50-day moving average
world_temp$mav50 <- as.numeric(stats::filter(world_temp$avg_temp,
                                             rep(1/50, 50), side = 1))
```

Same procedures were performed on the Victoria data:

```
#10-day moving average
victoria_temp$mav10 <- as.numeric(stats::filter(victoria_temp$avg_temp,
                                                rep(1/10, 10), side = 1))
#20-day moving average
victoria_temp$mav20 <- as.numeric(stats::filter(victoria_temp$avg_temp,
                                                rep(1/20, 20), side = 1))
#50-day moving average
victoria_temp$mav50 <- as.numeric(stats::filter(victoria_temp$avg_temp,
                                                rep(1/50, 50), side = 1))
```

Visualizations

Before visualization the data with lineplots, it was necessary to reshape the two data into long format. The combined dataset was also created. The `tidyverse` package includes some useful packages such as `ggplot2` and `dplyr` that provide simply solutions for data wrangling and visualizations.

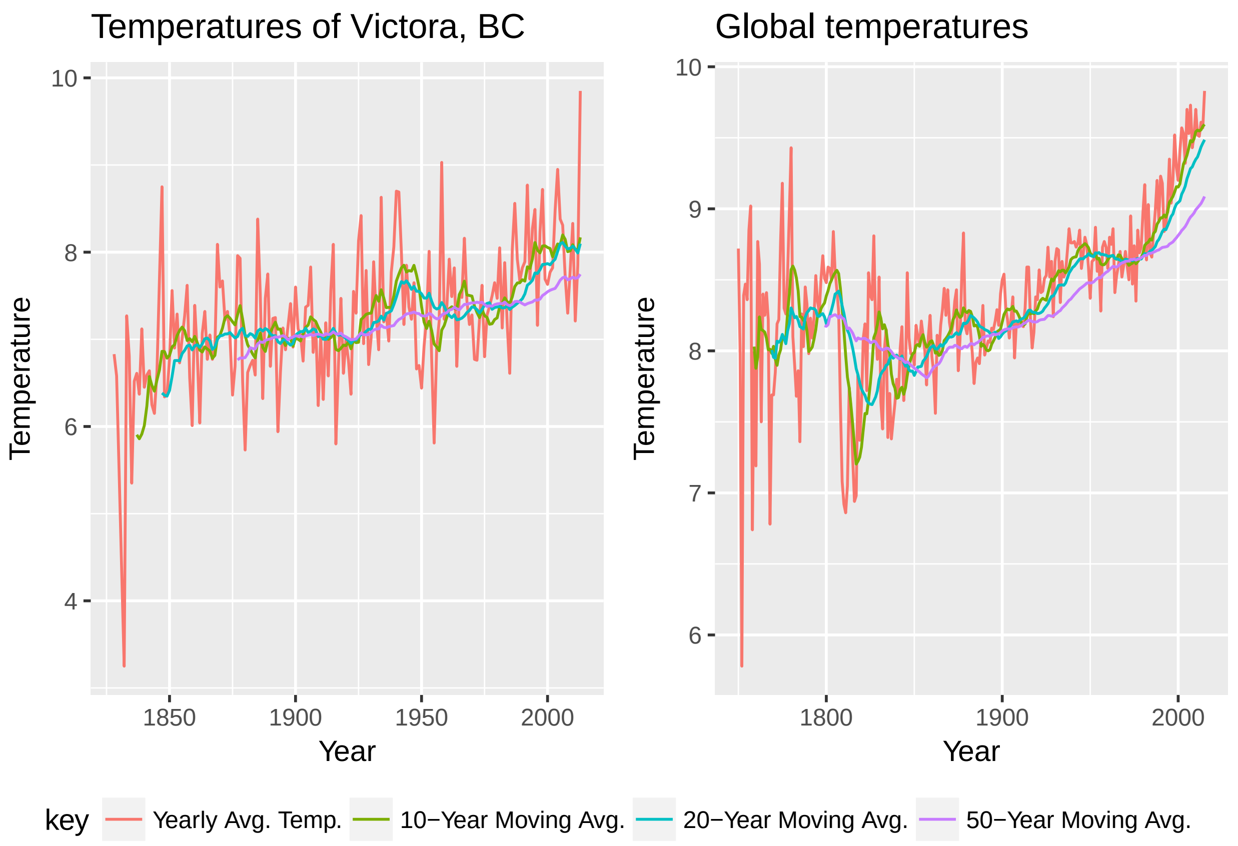
```
library(tidyverse)
#Remove the column 'country' from victoria_temp
victoria_temp <- select(victoria_temp, -country)
#Add column 'city' to world_temp
world_temp$city <- 'World'
world_temp <- select(world_temp, year, city, everything())
#combine data
combined <- rbind(world_temp, victoria_temp)

#transform to long format
victoria_temp <- gather(victoria_temp, key = 'key', value = 'temp', 3:6)
world_temp <- gather(world_temp, key = 'key', value = 'temp', 3:6)

#factorize 'key' to give proper names in plot legends
victoria_temp$key <- factor(victoria_temp$key,
                           levels = c('avg_temp', 'mav10', 'mav20', 'mav50'),
                           labels = c('Yearly Avg. Temp.',
                                      '10-Year Moving Avg.',
                                      '20-Year Moving Avg.',
                                      '50-Year Moving Avg.'))

world_temp$key <- factor(world_temp$key,
                        levels = c('avg_temp', 'mav10', 'mav20', 'mav50'),
                        labels = c('Yearly Avg. Temp.',
                                      '10-Year Moving Avg.',
                                      '20-Year Moving Avg.',
                                      '50-Year Moving Avg.'))
```

Looking the the lineplots separately for Victoria and global temperatures:



Observation 1

Both plots show strong fluctuations in yearly average temperatures (orange line), so moving averages are definitely better at shower the trend as they smooth out the fluctuations.

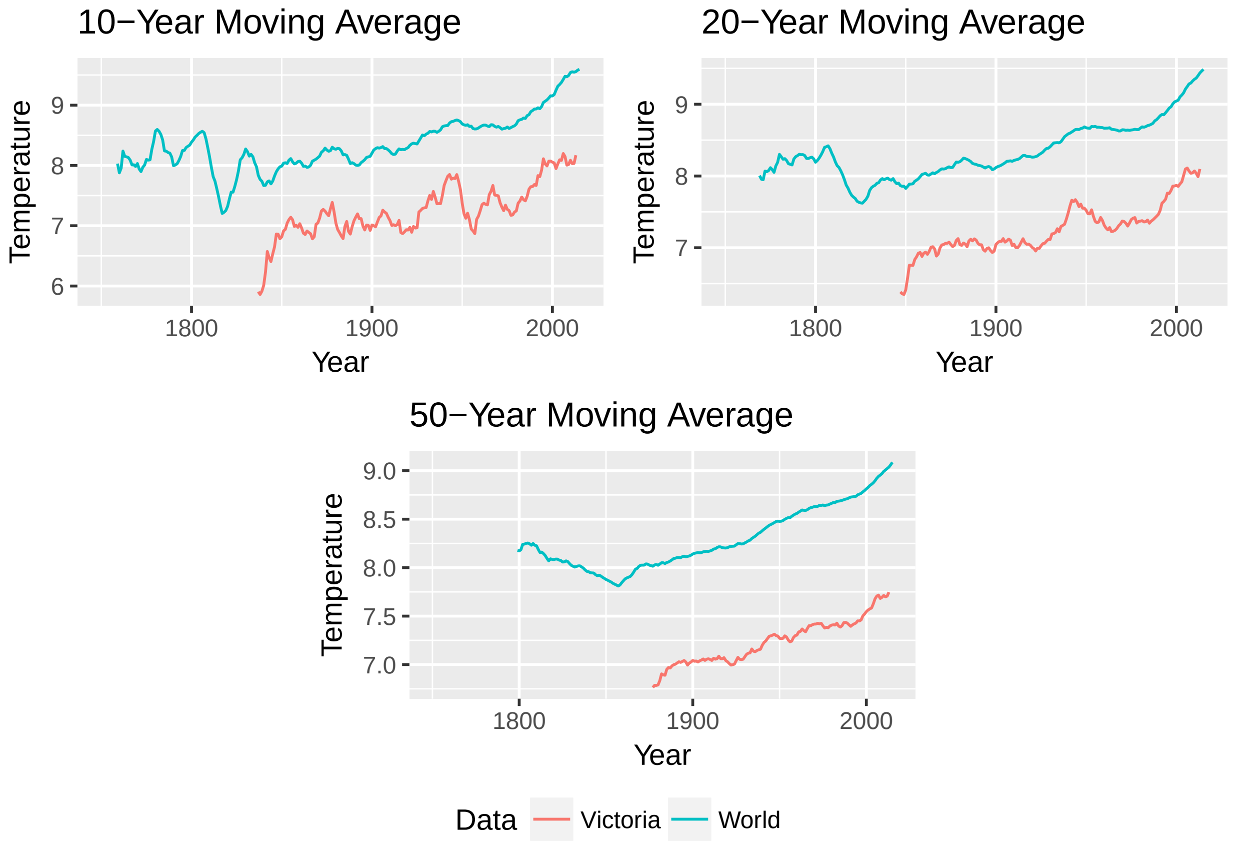
Observation 2

The yearly fluctuations in global temperatures are much more pronounced in the period of 1750 - 1850 compared to the mroe recent data. The yearly fluctuations in Victoral temperatures are relatively more stable.

Observation 3

Both plots suggest that, in the long run after smoothing out the fluctuations, there is a positive trend in average yearly temperature, which is a rather strong evidence in global warming or climate change.

The moving averages are plotted on the same graph to compare the relative change in temperatures between Victora and the world average.



Observation 4

All plots suggest that Victoria has relatively lower temperatures compared with the world's average.

Observation 5

Again, all short-, mid-, and long-term trends show the temperature has been increasing. The 50-year moving average provides a rather strong evidence that the increase has been quite steady. The world average temperature has increased by close to 2 degrees in the past 100 years.