# Exploring Weather Trends

Oscar Mendoza

9/3/2021

## Summary

In this Project, you will analyze local and global temperature data and compare the temperature trends where you live to overall global temperature trends.

## Instructions

Your goal will be to create a visualization and prepare a write up describing the similarities and differences between global temperature trends and temperature trend in the closest big city to where you live. To do this, you'll follow the steps belows:

- **Extract the data** from the database.
  - Write a SQL query to extract the city level data.
  - Write a SQL query to extract the global data.
- **Open up the CSV**
- **Create a line chart**
- **Make observations**
  - Is your city hotter or cooler on average compared to the global average? Has the difference been consistent over time?
  - "How do the changes in your city's temperature over time compare to the changes in the global average?"
  - What does the overall trend look like? Is the world getting hotter or cooler? Has the trend been consistent over the last few hundred year?

## Introduction

The present report was made using R and Rmarkdown, printed in a PDF File.

This report will have the following sections:

1. Data Acquisition and Processing
2. Data Visualization
3. Observations and Conclusions

## 1. Data Acquisition and Processing

The data is available in a Database Schema in Udacity's Project. There are 3 tables:

1. city_data

- year
- city
- country
- avg_temp

2. city_list

- city
- country

3. global_data

- year
- avg_temp

### City Selection

The city where I live is Santa Cruz de la Sierra, located in eastern Bolivia. Although there is some data regarding La Paz, which is the capital of Bolivia, this city is located in the Andes Plateau, with a climate different from the one of Santa Cruz. The grasslands are hotter and have an overall different behavior.

However, all the Brazilian cities that are in the database are in or nearby the coastline. No cities from Paraguay or were included, so due to proximity, La Paz will be analyzed.

```
SELECT *
FROM city_data
WHERE city = 'La Paz' AND country = 'Bolivia';
```

The earliest year when data was taken is 1855. By looking at the global_data table,

```
SELECT *
FROM global_data
ORDER BY year;
```

The earliest year is 1750. So in order to have the same time range, the years later than 1855 will be selected to 2013:

```
SELECT *
FROM global_data
WHERE global_data.year >= 1855 AND global_data.year <= 2013
ORDER BY year;
```

And the results will be saved into two different csv files:

- 'LaPaz_Weather.csv'
- 'Global_Weather.csv'

### Data Loading

The data will be loading into the following variables

```
global_data <- read.csv(file = "Global_Weather.csv", header = TRUE, sep = ",")

lapaz_data <- read.csv(file = "LaPaz_Weather.csv", header = TRUE, sep = ",")
```

The dimensions and structure of each table is:

```
#table sizes

dim(global_data)
```

```
## [1] 159   2
```

```
dim(lapaz_data)
```

```
## [1] 159   4
```

```
#table details

str(global_data)
```

```
## 'data.frame':    159 obs. of  2 variables:
##  $ year    : int  1855 1856 1857 1858 1859 1860 1861 1862 1863 1864 ...
##  $ avg_temp: num  8.11 8 7.76 8.1 8.25 7.96 7.85 7.56 8.11 7.98 ...
```

```
str(lapaz_data)
```

```
## 'data.frame':    159 obs. of  4 variables:
##  $ year    : int  1855 1856 1857 1858 1859 1860 1861 1862 1863 1864 ...
##  $ city    : chr  "La Paz" "La Paz" "La Paz" "La Paz" ...
##  $ country : chr  "Bolivia" "Bolivia" "Bolivia" "Bolivia" ...
##  $ avg_temp: num  7.4 7.81 7.66 8.02 8.7 8.41 8.05 7.96 8.27 8.05 ...
```

### Data Processing

Once the content of the tables have been described, a new column is added: mavg, moving average. In this case, we will use a n=5, which means that the moving average will be computed for a number of years, defined by the variable $n$ which is a numerical vector.

```
#compute the moving average for a number of time ranges

n <- c(2, 5, 10, 20, 25, 50)

#moving averages for La Paz city

for (i in 1:length(n)){
  column_name <- paste("mavg", n[i], sep = "")
```

```
  new_data <- rep(0, nrow(lapaz_data))
  for (j in 1:(nrow(lapaz_data)-n[i]+1)){
    new_data[j+n[i]-1] <- sum(lapaz_data$avg_temp[j:(j+n[i]-1)])/n[i]
  }
  lapaz_data[ , ncol(lapaz_data) + 1] <- new_data
  colnames(lapaz_data)[ncol(lapaz_data)] <- column_name
}


#moving average for global data

for (i in 1:length(n)){
  column_name <- paste("mavg", n[i], sep = "")
  new_data <- rep(0, nrow(global_data))
  for (j in 1:(nrow(global_data)-n[i]+1)){
    new_data[j+n[i]-1] <- sum(global_data$avg_temp[j:(j+n[i]-1)])/n[i]
  }
  global_data[ , ncol(global_data) + 1] <- new_data
  colnames(global_data)[ncol(global_data)] <- column_name
}
```

And then we check the consistency of the computation

```
head(lapaz_data)
```

```
##   year   city country avg_temp mavg2 mavg5 mavg10 mavg20 mavg25 mavg50
## 1 1855 La Paz Bolivia     7.40 0.000 0.000      0      0      0      0
## 2 1856 La Paz Bolivia     7.81 7.605 0.000      0      0      0      0
## 3 1857 La Paz Bolivia     7.66 7.735 0.000      0      0      0      0
## 4 1858 La Paz Bolivia     8.02 7.840 0.000      0      0      0      0
## 5 1859 La Paz Bolivia     8.70 8.360 7.918      0      0      0      0
## 6 1860 La Paz Bolivia     8.41 8.555 8.120      0      0      0      0
```

```
head(global_data)
```

```
##   year avg_temp mavg2 mavg5 mavg10 mavg20 mavg25 mavg50
## 1 1855     8.11 0.000 0.000      0      0      0      0
## 2 1856     8.00 8.055 0.000      0      0      0      0
## 3 1857     7.76 7.880 0.000      0      0      0      0
## 4 1858     8.10 7.930 0.000      0      0      0      0
## 5 1859     8.25 8.175 8.044      0      0      0      0
## 6 1860     7.96 8.105 8.014      0      0      0      0
```

A third dataframe is built with all the measurements in order to create a plot to compare them

```
#a new dataframe is created, with the following  columns: year, source,
#variable, value

library(reshape2)

lapaz_computeddata <- melt(lapaz_data, id = c("year",
```

```
                                            "city",
                                            "country"))

#rename the column names
names(lapaz_computeddata) <- c("year",
                               "source",
                               "country",
                               "variable",
                               "value")

#drop the column "country" because is no longer necessary

lapaz_computeddata <- subset(lapaz_computeddata, select = c("year",
                                                            "source",
                                                            "variable",
                                                            "value"))

#we do the same for the global dataframe

global_computeddata <- melt(global_data, id = c("year"))

#add the column "source

global_computeddata$source <- "Global"

#reorder the DataFrame

global_computeddata <- global_computeddata[ , c(1, 4, 2, 3)]

#bind both dataframes

alldata <- rbind(lapaz_computeddata, global_computeddata)
```

## 2. Data Visualization

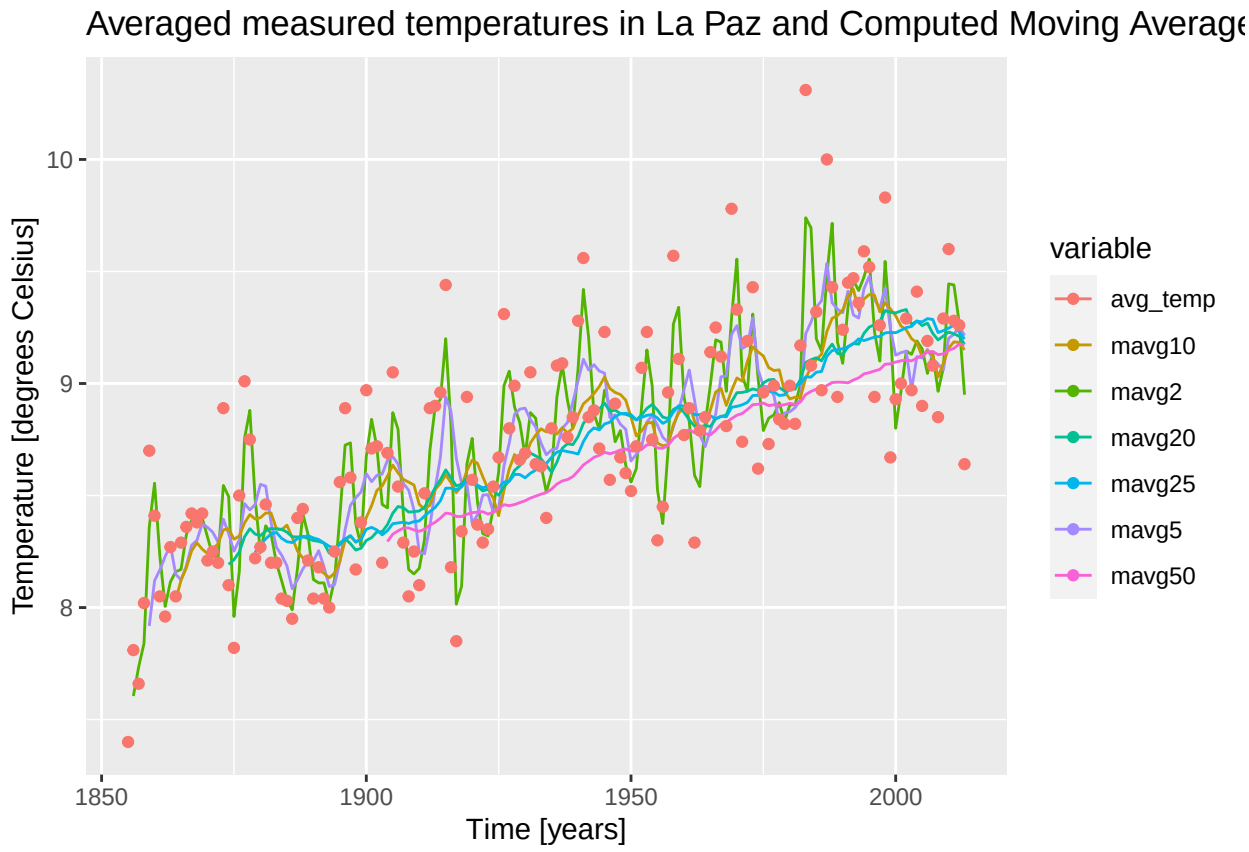
The ggplot2 package will be used for this visualization

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

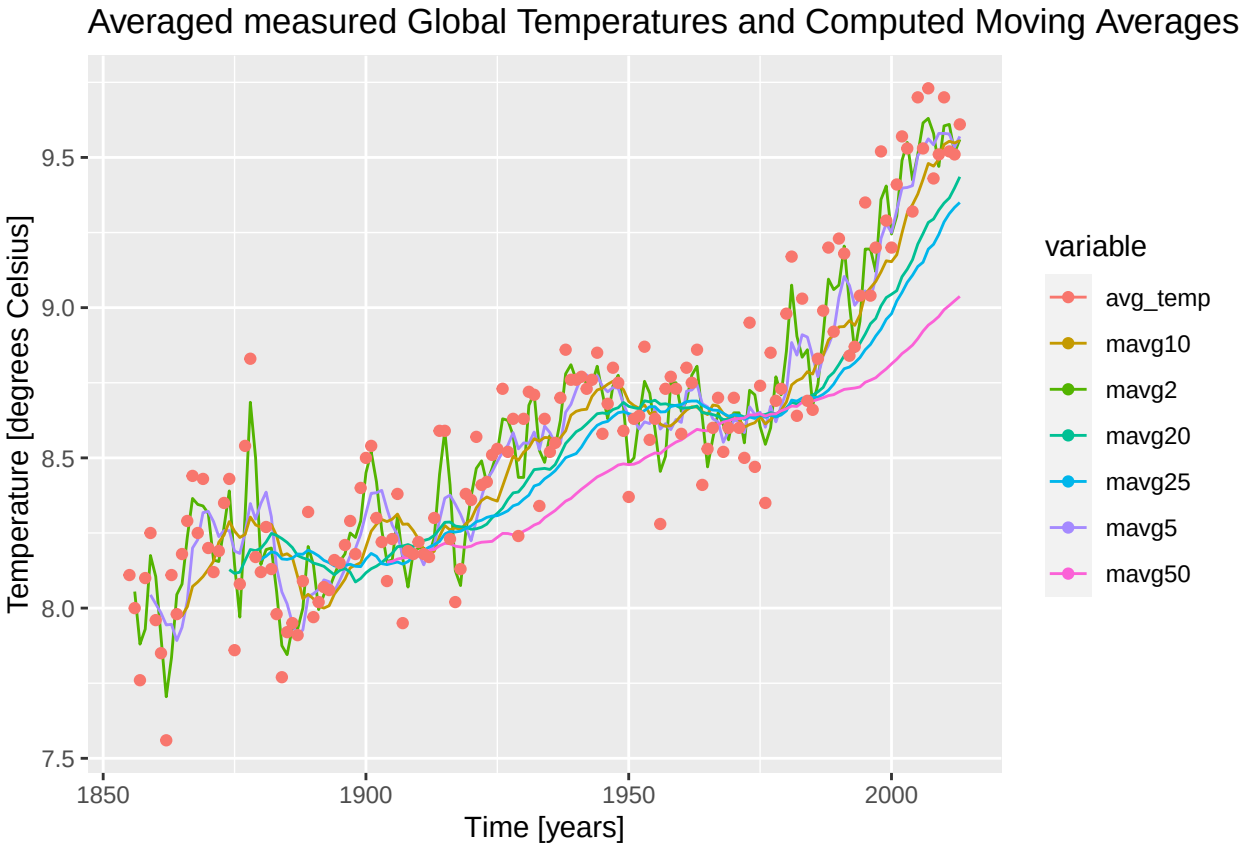
So the data in La Paz will be reviewed first

```
lapaz_avg_temp <- lapaz_computeddata %>%
  filter(variable == "avg_temp", value > 0)

lapaz_computeddata %>%
  filter(variable != "avg_temp", value > 0) %>%
  ggplot (aes(x = year, y = value, color = variable)) +
  geom_line() +
  geom_point(data = lapaz_avg_temp, aes(x = year, y = value)) +
  ggtitle("Averaged measured temperatures in La Paz and Computed Moving Averages") +
  xlab("Time [years]") +
  ylab("Temperature [degrees Celsius]")
```




And then the Global data

```
global_avg_temp <- global_computeddata %>%
  filter(variable == "avg_temp", value > 0)

global_computeddata %>% filter(variable != "avg_temp", value > 0) %>%
  ggplot(aes(x = year, y = value, color = variable)) +
  geom_line() +
  geom_point(data = global_avg_temp, aes(x = year, y = value)) +
  ggtitle("Averaged measured Global Temperatures and Computed Moving Averages") +
  xlab("Time [years]") +
  ylab("Temperature [degrees Celsius]")
```

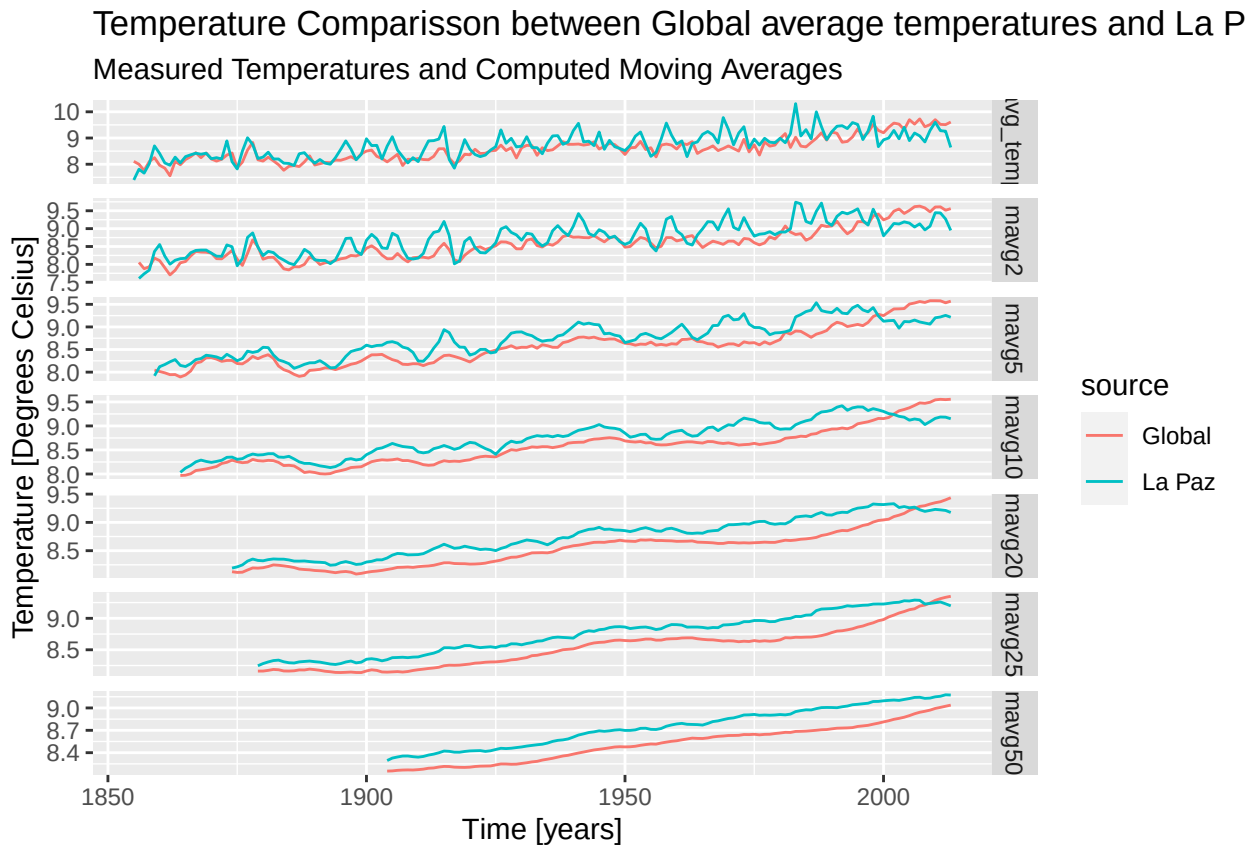



## 3. Observations and Conclusions

- **Observation 1**

  By looking at the previous plots it is clear that both plots, La Paz and Global, show an increasing trend. Of all the moving averages computed, the one that showed the most improved trend visualization is the moving average with a time range of 25 years. On the other hand, by looking at the moving average with a time range of 50 years, the trend is oversmoothed.

```
alldata %>%
  filter(value > 0) %>%
  ggplot( aes(x = year, y =value, color = source)) +
  geom_line() +
  facet_grid(variable~., scales = "free") +
  ggtitle("Temperature Comparisson between Global average temperatures and La Paz",
          subtitle = "Measured Temperatures and Computed Moving Averages") +
  xlab("Time [years]") +
  ylab("Temperature [Degrees Celsius]")
```




- **Observation 2**

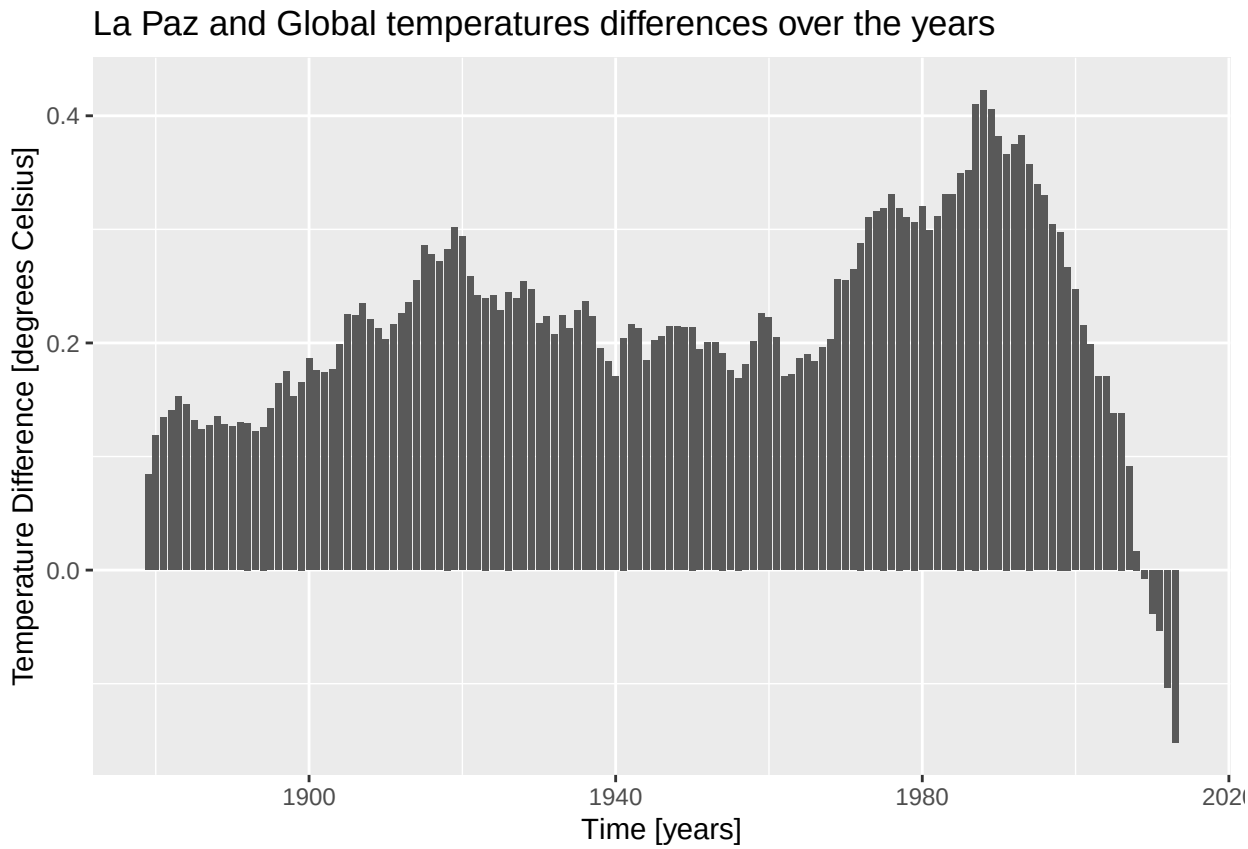
  It is clear that, in all cases, the general trend shows an increment in the average temperature over the years, in both the Global temperatures and in La Paz.

### Questions

So for answering the following questions, the computed moving average of 25 years will be used.

- Is your city hotter or cooler on average compared to the global average? Has the difference been consistent over time?

```
#a new dataframe is built
mavg25 <- data.frame(year = lapaz_data$year, lapaz_mavg25 = lapaz_data$mavg25,
                     global_mavg25 = global_data$mavg25)

#the difference is defined as La Paz - Global mvag25 temperatures
mavg25$difference <- mavg25$lapaz_mavg25 - mavg25$global_mavg25

mavg25 %>%
  filter(lapaz_mavg25 > 0, global_mavg25 > 0) %>%
  ggplot(aes(x = year, y = difference)) +
  geom_col() +
  ggtitle("La Paz and Global temperatures differences over the years") +
  xlab("Time [years]") +
  ylab("Temperature Difference [degrees Celsius]")
```




*Answer*: The column plot show the difference between La Paz's average temperature from the computed mavg25. Although the weather in La Paz has been consistenly hotter than the global average, the trend has changed over the last few years.

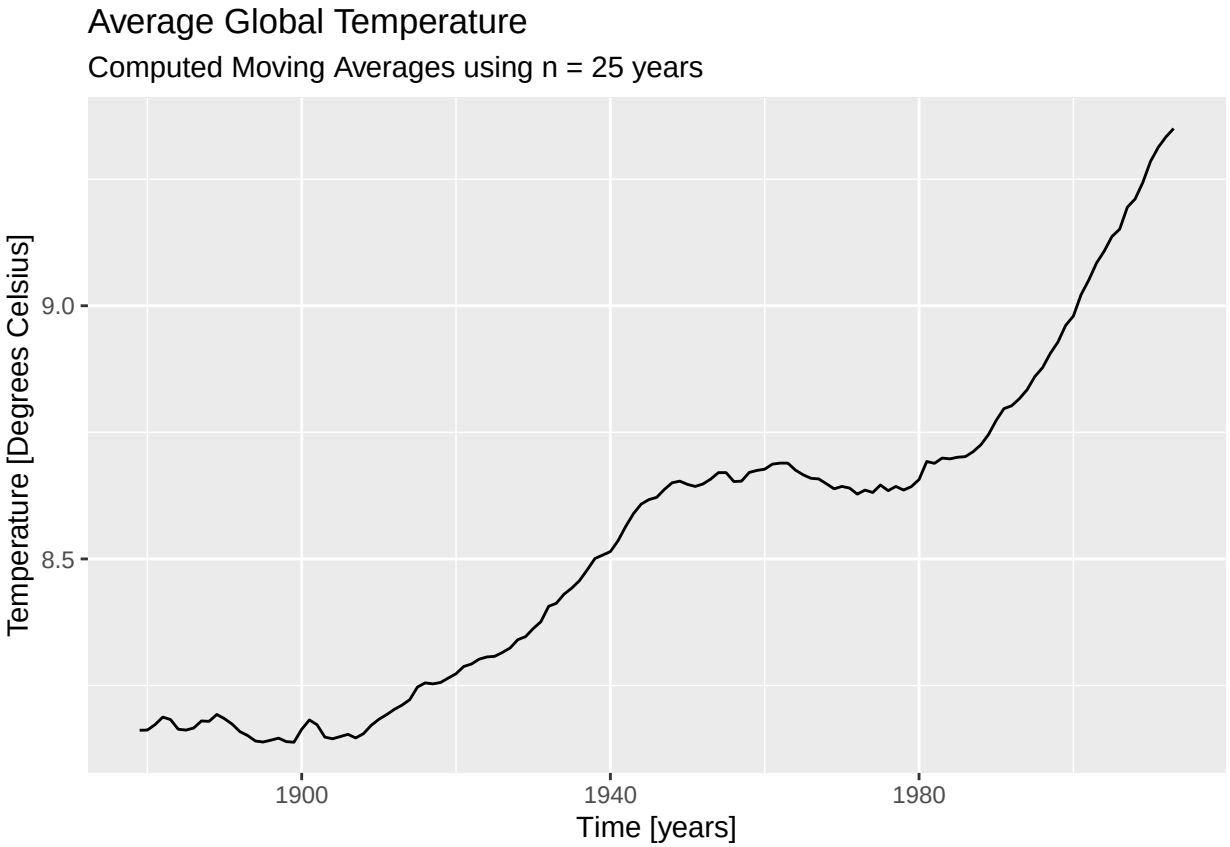
- "How do the changes in your city's temperature over time compare to the changes in the global average?"

  *Answer*: The follow the same trend, an increment in the average temperature have been observed.

- What does the overall trend look like? Is the world getting hotter or cooler? Has the trend been consistent over the last few hundred year?

  *Answer*: The world has been getting hotter, and this trend it's been constant for the last few hundred years.

```
alldata %>% filter(value > 0, variable == "mavg25", source == "Global") %>%
  ggplot( aes(x = year, y =value)) +
  geom_line() +
  ggtitle("Average Global Temperature",
          subtitle = "Computed Moving Averages using n = 25 years") +
  xlab("Time [years]") +
  ylab("Temperature [Degrees Celsius]")
```




Although, if we check the temperature after 1980, we can see that the average temperature is increasing faster than in previous years.