




FIRST DAND PROJECT



Chrysanthi Polyzoni

Outline

In order to correctly extract data from tables provided two simple SQL queries have been used

1. City Data have been extracted for the City of Milan in Italy and the query used is provided here below:

```
SELECT * FROM city_data
WHERE city_data.country = 'Italy'
```

2. Global Data have been extracted using a very simple general query:

```
SELECT * FROM global_data
```

The data was saved as .csv files and saved locally at C:\\Users\\User\\Documents\\SQL_climate_data separately as 'global_raw_data.csv' and 'milan_raw_data.csv' for Global Data and Milan data. It has been decided to perform separate analysis and not to join data since tables provided different table dimensions, hence raw data could provide meaningful information.

Coma Separated Values files aka .csv files are easily imported and manipulated using R's built-in "utils" package read.csv() command in R Studio as follows

```
global_raw_data <- read.csv('C:\\Users\\User\\Documents\\SQL_climate_data\\global_raw_data.csv')
milan_raw_data <- read.csv('C:\\Users\\User\\Documents\\SQL_climate_data\\milan_raw_data.csv')
```

Using dim() we confirmed the difference of our datasets

```
dim(global_raw_data)
dim(milan_raw_data)
```

global_raw_data is made out of 266 rows and 2 columns while
milan_raw_data is made out of 542 rows and four columns

Some summary statistics for both datasets have been consulted

```
summary(global_raw_data$avg_temp)
summary(milan_raw_data$avg_temp)

> summary(global_raw_data$avg_temp)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.780   8.082   8.375   8.369   8.707   9.830
> summary(milan_raw_data$avg_temp)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
  0.180   6.790   8.395   9.354  11.928  13.620      8
```

A couple of plots demonstrating raw data are illustrated in Figure 1 and Figure 2

Using the following commands

```
plot(x = milan_raw_data$year, y = milan_raw_data$avg_temp, main = "Milan Temperature", xlab =
"Years", ylab = "Degrees in Celcius")
```

```
plot(x = global_raw_data$year, y = global_raw_data$avg_temp, main = "Global Temperature", xlab =
"Years", ylab = "Degrees in Celcius")
```

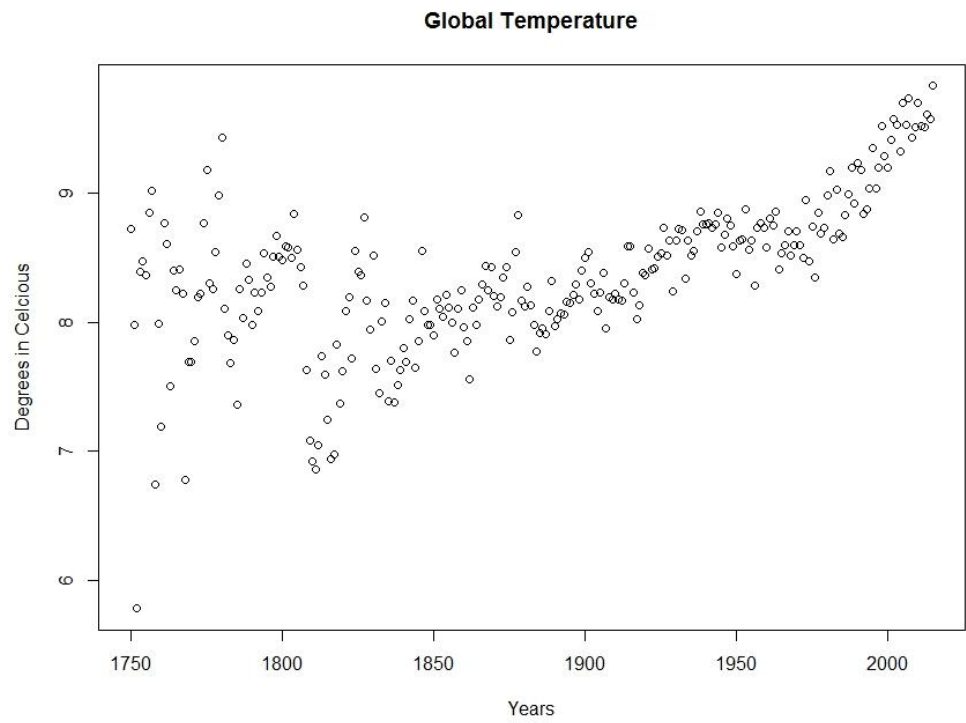


Figure 1

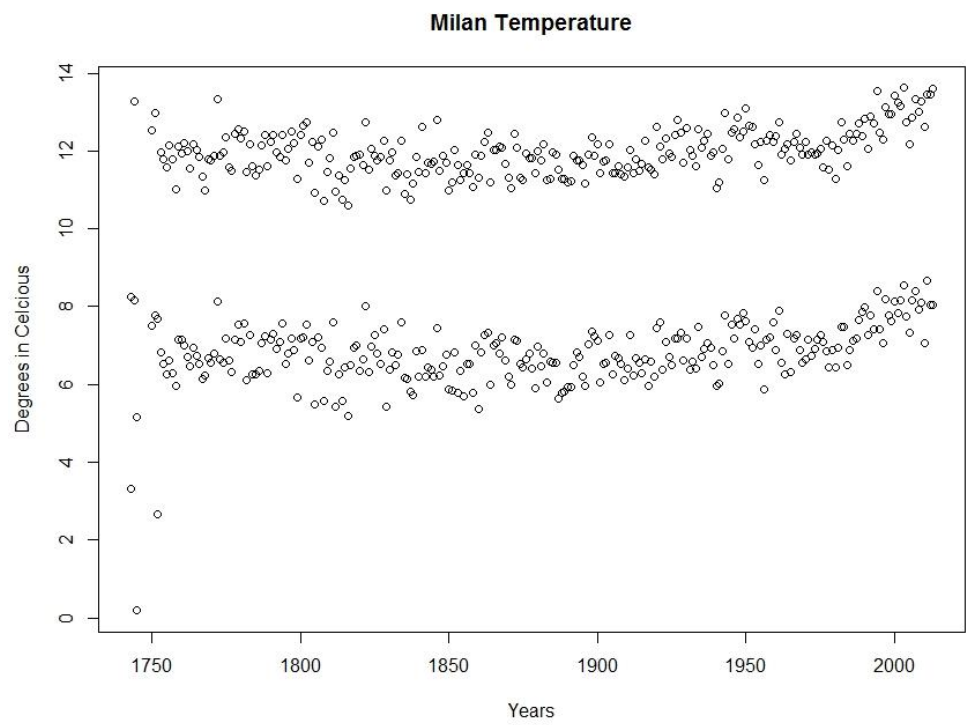


Figure 2

Plots have then been removed using `graphics.off()` command.

Line Charts

In order to make some sense of our data we have calculated moving averages after installing forecast package so as to use `ma()` function – moving average function

```
install.packages("forecast")
```

```
library(forecast)
```

Here is how 30-year moving averages for Milan data and Global data have been calculated

```
global_moving_average <- ma(global_raw_data$avg_temp, order = 30, centre = TRUE)
```

```
milan_moving_average <- ma(milan_raw_data$avg_temp, order = 30, centre = TRUE)
```

and plotted in Figure 3 representing moving average for Global Average Temperature and Figure 4 representing moving average for Milan temperature.

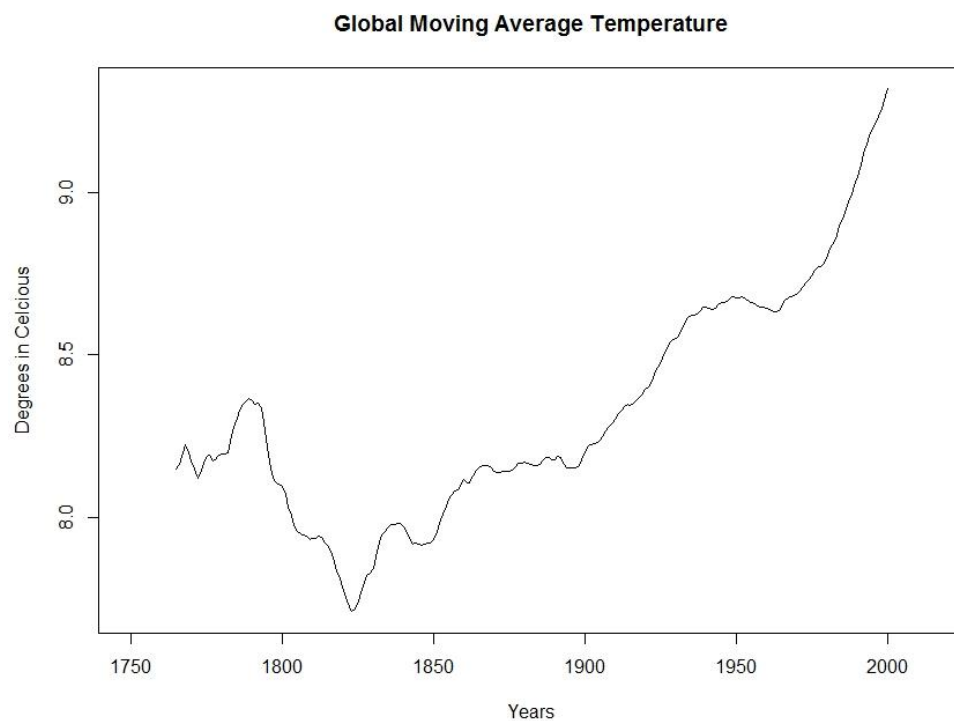


Figure 3

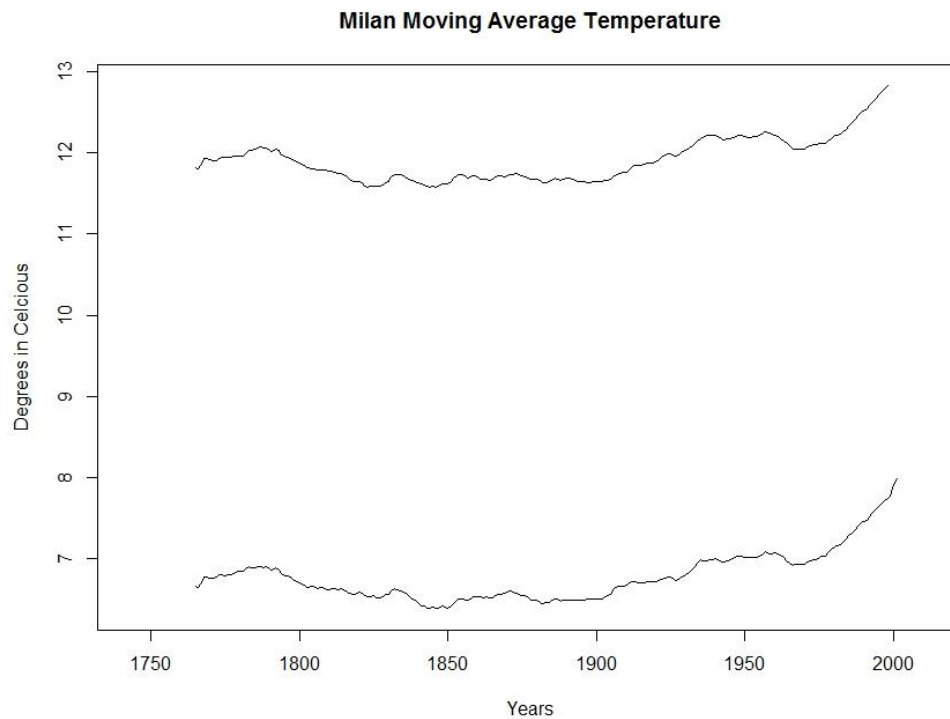


Figure 4

Observations

- We can observe from Figure 1 that global raw temperature data have a higher variance before the year 1900 with respect to years after 1900.
- We can observe from Figure 2 that local datasets include seasonal information that are aggregated globally because winter and summer temperatures eliminate one the other because they include separate temperature for the two earths hemispheres.
- We can observe from Figure 3 that after the year 1830 global temperature started to increase
- We can observe from Figure 4 that Milan's temperature saw very little overall increase with respect to global data.