PROJECT: EXPLORE WEATHER TRENDS

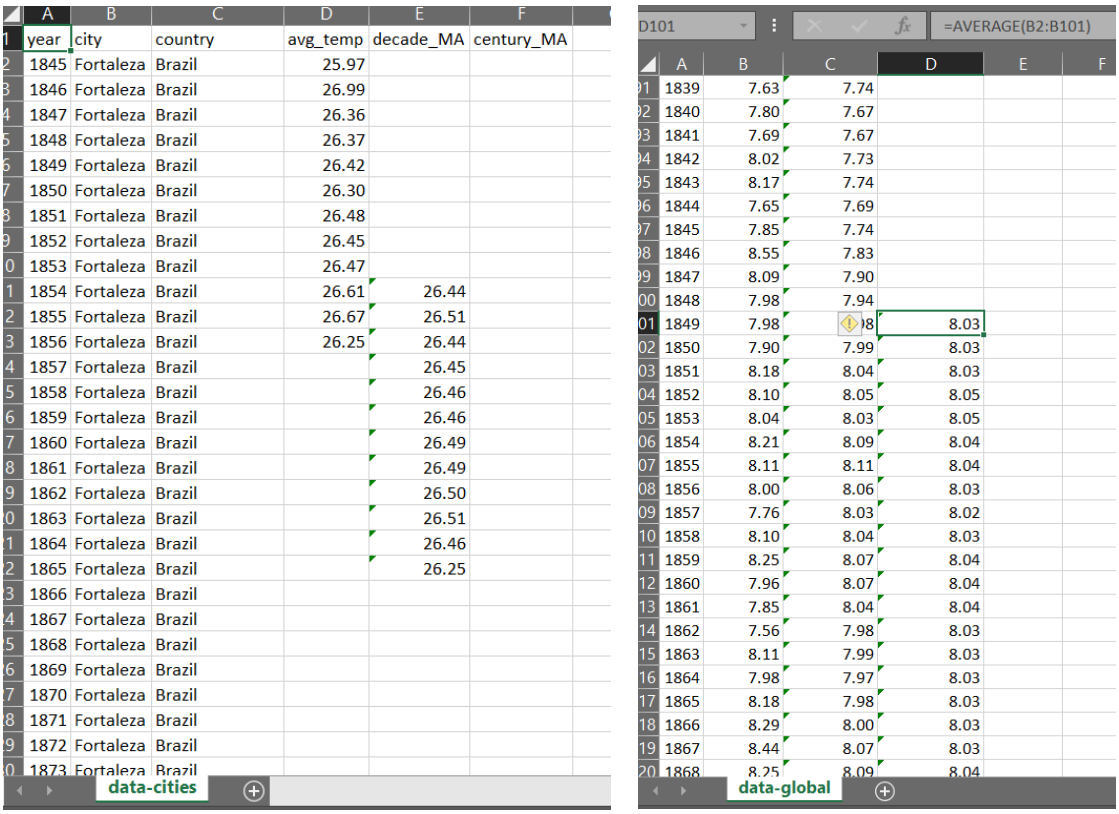
EXTRACT DATA

Below, is shown the blocks of code (highlighted in `Consolas` font) I used to extract weather information and data using SQL. First, I checked all available cities in the database (item 1). I chose Munich first, because it seemed to be the city nearest to where I live today. Additionally, I chose other cities like Seattle (where some cool bands I like came from) and Fortaleza (where I was born), but also other famous cities, such as London and Paris. Specifically for London, I noticed that there are actually two (one in Canada and other in UK). Since I wanted to get data only for the one in UK, I split condition to correctly select the one I chose (item 2). Finally, I got the entire available data for the global weather, given that no special subset was declared in the project description (item 3).

```
/* 1- Query to obtain information on available cities that are close to where I live*/

SELECT *

FROM city_list;


/* 2- Query to obtain weather data on those cities that will be used in the project*/

SELECT *

FROM city_data

WHERE city IN ('Munich','Fortaleza','Seattle','Paris') OR (city LIKE 'London' AND country NOT LIKE
'Canada');


/* 3- Query to obtain global temperature*/

SELECT *

FROM global_data;
```

MOVING AVERAGES

To calculate the moving averages, I loaded the csv files in Excel and produced the averages using the formula **=AVERAGE()**, as explained in the course. Since we are talking about years, it makes more sense to see them as decades and centuries, in addition to the individual years. Therefore, I produced two columns on the csv files for moving averages, e.g. for 10-year- and 100-year-intervals (Fig. 1).



year	city	country	avg_temp	decade_MA	century_MA
1845	Fortaleza	Brazil	25.97		
1846	Fortaleza	Brazil	26.99		
1847	Fortaleza	Brazil	26.36		
1848	Fortaleza	Brazil	26.37		
1849	Fortaleza	Brazil	26.42		
1850	Fortaleza	Brazil	26.30		
1851	Fortaleza	Brazil	26.48		
1852	Fortaleza	Brazil	26.45		
1853	Fortaleza	Brazil	26.47		
1854	Fortaleza	Brazil	26.61	26.44	
1855	Fortaleza	Brazil	26.67	26.51	
1856	Fortaleza	Brazil	26.25	26.44	
1857	Fortaleza	Brazil		26.45	
1858	Fortaleza	Brazil		26.46	
1859	Fortaleza	Brazil		26.46	
1860	Fortaleza	Brazil		26.49	
1861	Fortaleza	Brazil		26.49	
1862	Fortaleza	Brazil		26.50	
1863	Fortaleza	Brazil		26.51	
1864	Fortaleza	Brazil		26.46	
1865	Fortaleza	Brazil		26.25	
1866	Fortaleza	Brazil			
1867	Fortaleza	Brazil			
1868	Fortaleza	Brazil			
1869	Fortaleza	Brazil			
1870	Fortaleza	Brazil			
1871	Fortaleza	Brazil			
1872	Fortaleza	Brazil			
1873	Fortaleza	Brazil			

data-cities

D101 =AVERAGE(B2:B101)

1839	7.63	7.74	
1840	7.80	7.67	
1841	7.69	7.67	
1842	8.02	7.73	
1843	8.17	7.74	
1844	7.65	7.69	
1845	7.85	7.74	
1846	8.55	7.83	
1847	8.09	7.90	
1848	7.98	7.94	
1849	7.98	8	8.03
1850	7.90	7.99	8.03
1851	8.18	8.04	8.03
1852	8.10	8.05	8.05
1853	8.04	8.03	8.05
1854	8.21	8.09	8.04
1855	8.11	8.11	8.04
1856	8.00	8.06	8.03
1857	7.76	8.03	8.02
1858	8.10	8.04	8.03
1859	8.25	8.07	8.04
1860	7.96	8.07	8.04
1861	7.85	8.04	8.04
1862	7.56	7.98	8.03
1863	8.11	7.99	8.03
1864	7.98	7.97	8.03
1865	8.18	7.98	8.03
1866	8.29	8.00	8.03
1867	8.44	8.07	8.03
1868	8.25	8.09	8.04

data-global

Figure 1. Csv file overviews for the cities (left) and global temperature data (right).

DATA VISUALIZATION

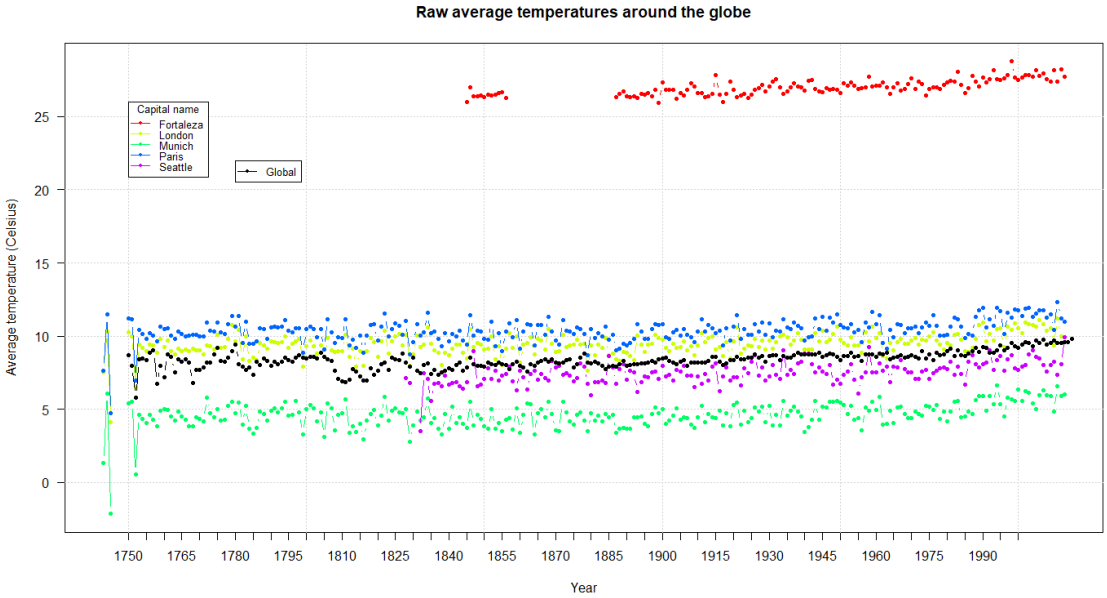
For data visualization and analyses I used R, because I feel myself more comfortable in that particular environment, and also because I understand the language and know about functions that could be used to solve this project.

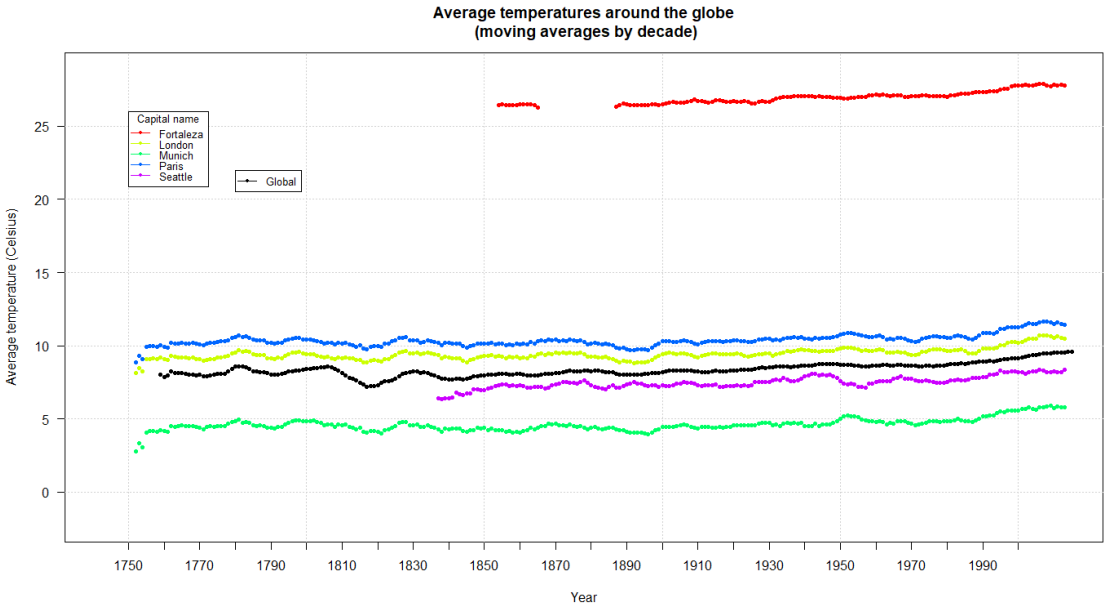
First I loaded the two weather datasets (for cities and global). Then I used a simple plot function to plot the three forms of temperature data (e.g. raw data by year, moving average by decade, and moving average by century). I used google to get information on codes necessary for plotting lines with colors and adding legends to the graph. The information was taken from the following link: https://www.statmethods.net/graphs/line.html. I believe it did not influence the core aim of learning in this project. For clarity, I include below the R code, together with the graphical output.

```
## load datasets
cities <- read.csv("data-cities.csv", sep=";")
global <- read.csv("data-global.csv", sep=";")

# Raw average mean temperatures ----------------------------------------
## convert factor to numeric for convenience
cities$city2 <- as.numeric(cities$city)
ncities <- max(cities$city2)

# get the range for the x and y axis
xrange <- range(cities$year)
yrange <- range(cities$avg_temp, na.rm = T)

## set up the plot
plot(xrange, yrange,
     type="n",
     xlab="Year",
     ylab="Average temperature (Celsius)",
     axes=FALSE
     )

ticks = seq(1750,2000,5)
axis(side = 1, at = ticks)
axis(side = 2, las=1)
box()
grid()

## set colors
colors <- rainbow(ncities)

## add lines from particular cities
for (i in 1:ncities) {
  city <- subset(cities, city2==i)
  lines(city$year, city$avg_temp,
        type="b", lwd=1.5,
        pch=20, lty=1, col=colors[i])
  }

## add line for global trend
lines(global$year, global$avg_temp,
      type="b", lwd=1.5,
      pch=20, lty=1, col="black"
      )

## add a title and subtitle
title("Raw average temperatures around the globe")

## add legends
levels(cities$city)
legend(1750, 26,
```

```
       levels(cities$city), cex=0.8, col=colors,
       pch=20, lty=1, title="Capital name"
       )

legend(1780, 22,
       "Global", cex=0.8, col="black",
       pch=20, lty=1
       )


# Mean temperature moving averages by decade  -----------------------------
## set up the plot
plot(xrange, yrange,
     type="n",
     xlab="Year",
     ylab="Average temperature (Celsius)",
     axes=FALSE
)

ticks = seq(1750,2000,10) #for decades
axis(side = 1, at = ticks)
axis(side = 2, las=1)
box()
grid()

## add lines from particular cities
for (i in 1:ncities) {
  city <- subset(cities, city2==i)
  lines(city$year, city$decade_MA,
        type="b", lwd=1.5,
        pch=20, lty=1, col=colors[i])
}

## add line for global trend
lines(global$year, global$decade_MA,
      type="b", lwd=1.5,
      pch=20, lty=1, col="black"
)

## add a title and subtitle
title("Average temperatures around the globe\n (moving averages by decade)")

## add legends
legend(1750, 26,
       levels(cities$city), cex=0.8, col=colors,
       pch=20, lty=1, title="Capital name"
)

legend(1780, 22,
       "Global", cex=0.8, col="black",
       pch=20, lty=1
)


# Mean temperature moving averages by century -----------------------------
## set up the plot
plot(xrange, yrange,
     type="n",
     xlab="Year",
     ylab="Average temperature (Celsius)",
     axes=FALSE
)

ticks = seq(1750,2000,50) #for centuries: longer interval
axis(side = 1, at = ticks)
axis(side = 2, las=1)
box()
grid()

## add lines from particular cities
```

```
for (i in 1:ncities) {
  city <- subset(cities, city2==i)
  lines(city$year, city$century_MA,
        type="b", lwd=1.5,
        pch=20, lty=1, col=colors[i])
}

## add line for global trend
lines(global$year, global$century_MA,
      type="b", lwd=1.5,
      pch=20, lty=1, col="black"
)

## add a title and subtitle
title("Average temperatures around the globe\n (moving averages by decade)")

## add legends
legend(1750, 26,
       levels(cities$city), cex=0.8, col=colors,
       pch=20, lty=1, title="Capital name"
)

legend(1780, 22,
       "Global", cex=0.8, col="black",
       pch=20, lty=1
)
```





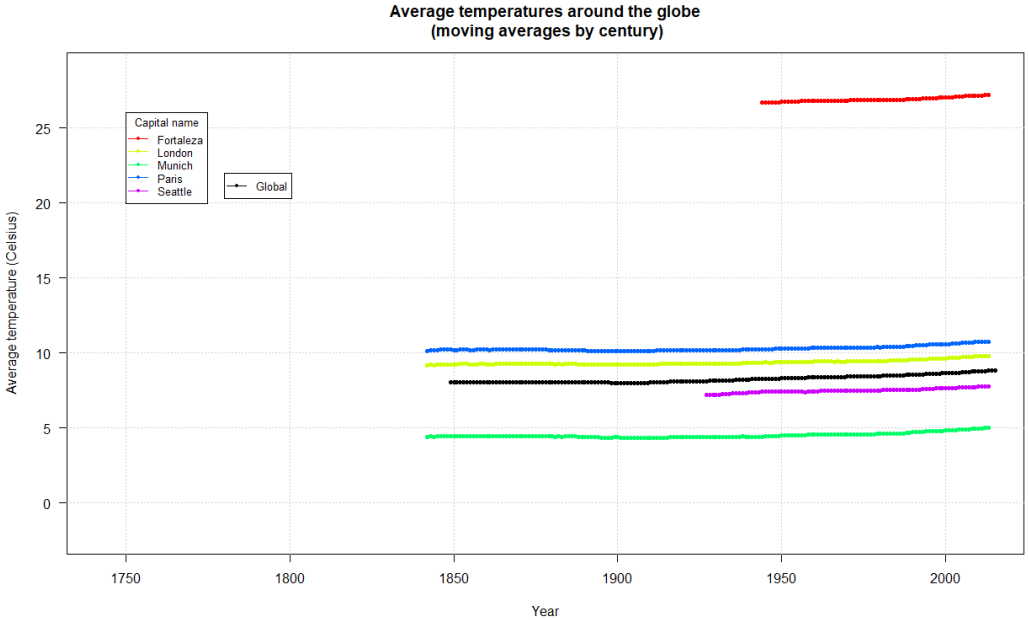


Figure 2. Global and local average temperatures for the raw data and moving averages used in this project.

OBSERVATIONS

There are a couple of observations that we can take from the data and the graphs shown here.

Are temperatures constant or increasing?

Visually it seems that overall temperatures are constant, because lines are more or less flat in the graphs (Fig. 2). Using the raw data we can see that there is a lot of fluctuation like a fine noise. When we smooth the information into longer intervals of years, the lines get more straight, helping visualize the overall trend.

When we use a simple generalized linear model to test the statistical significance of the effect of year on average temperature, however, we get another result. Global temperatures are in average indeed increasing by a very small estimate, as show below. Of course to our eyes it seems to small of an effect, but in terms of temperature in Celsius it actually may be a big deal. If we look with attention into Fig. 1, actually seems that lines go up, especially towards the final years.

```
summary(lm(avg_temp ~ year, data = global))
```

```
Call: lm(formula = avg_temp ~ year, data = global)

Residuals:
     Min       1Q   Median       3Q      Max
-1.97174 -0.26364 -0.04193  0.28744  1.54572

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.541564   0.689582  -0.785    0.433
year         0.004734   0.000366  12.933   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4584 on 264 degrees of freedom
Multiple R-squared:  0.3878,    Adjusted R-squared:  0.3855
F-statistic: 167.3 on 1 and 264 DF,  p-value: < 2.2e-16
```

Are there differences between cities regarding average temperatures?

The striking difference seen in Fig. 2 is the temperature reported for Fortaleza. It contrasts with all other cities, including the global temperature. This is because Fortaleza is a tropical city, whereas the other cities are located in temperate zones, and thus temperatures are much lower, close to the 10 °C. Using an extension of the R formula above to test the significance of these differences, I also found that they are statistically supported.

```
model <- glm(avg_temp ~ year + city, data = model)
summary(model)
library(multcomp)
glht(model, mcp(city = "Tukey"))
```

```
Call:glm(formula = avg_temp ~ year + city, data = model)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.63060  -0.37445  -0.00707   0.38213   1.80072
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.231e+01  9.666e-01   12.73   <2e-16 ***
year         7.588e-03  4.873e-04   15.57   <2e-16 ***
cityLondon  -1.736e+01  8.419e-02 -206.14   <2e-16 ***
cityMunich  -2.220e+01  8.419e-02 -263.69   <2e-16 ***
cityParis   -1.643e+01  8.419e-02 -195.15   <2e-16 ***
citySeattle -1.943e+01  9.120e-02 -213.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.321942)

    Null deviance: 26526.62  on 672  degrees of freedom
Residual deviance:   214.74  on 667  degrees of freedom
AIC: 1155.1

Number of Fisher Scoring iterations: 2
```

And boxplots further show these differences (Fig. 3).

```
par(mfrow = c(1, 2))
plot(avg_temp ~ city,  ylim = c(0,30), data=cities,
     xlab = "Capital city", ylab = "Average temperature (Celsius)")
boxplot(global$avg_temp, ylim = c(0,30),
        xlab = "Globe", ylab = "Average temperature (Celsius)")
```

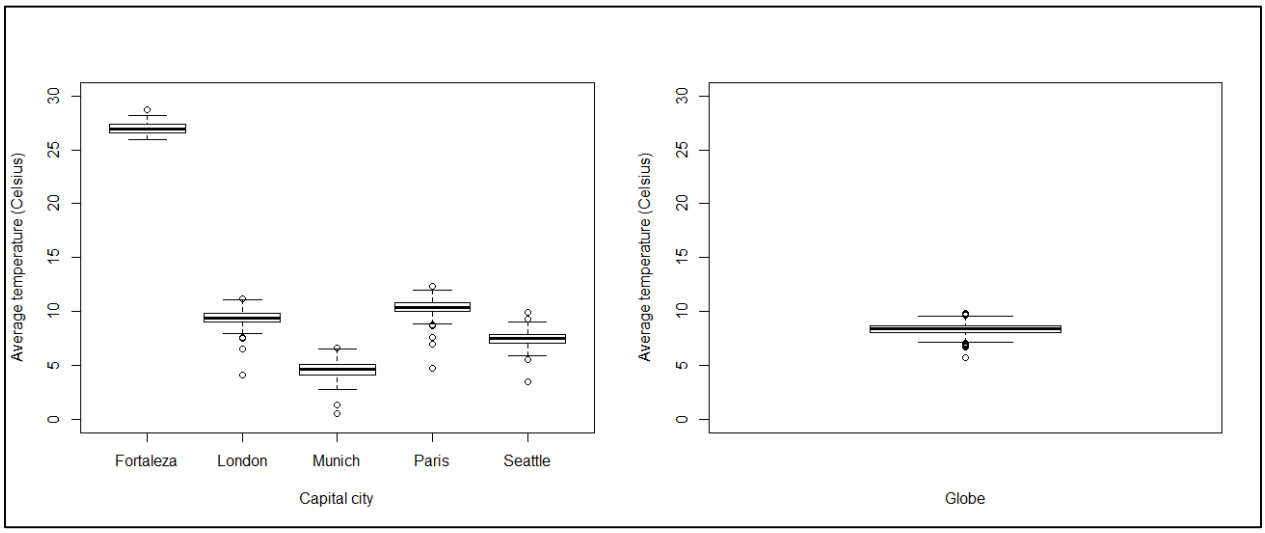


Figure 3. Boxplots showing mean average temperature for the five capital cities used in the project (left) and the global temperature data (right). Note that Fortaleza has a mean temperature over 25 °C, whereas all other are close to 10 °C.

Missing values

One issue I noticed is the presence of missing values. These are clearly seen in the first pane of Fig. 2. Missing data was specially found in early records, where probably data observation was harder to take or there were no instruments to do this at that time. As a consequence, moving averages cannot be calculated in early years. This brings to another important point that, in order to calculate moving averages and smooth the trend into straighter lines, we also loose fine information on the data, not exactly getting

averages from the “beginning”. The important question is whether this influences your goal of the analysis, and it should be taken into consideration. For an overall trend, moving averages seems a good option.

Are there differences between my city and the global temperatures?

Based on Fig. 2, we can observe that Munich (green line) is cooler than the average (black line).

After performing a simple T-Test between the unpaired samples observed for the global temperatures and Munich, I found statistical support that in my city temperatures are significantly lower than as for the globe. This can be further visualized in Fig. 3.

```
MUNICH <- subset(cities, city=="Munich")
t.test(MUNICH$avg_temp, global$avg_temp,
       alternative = c("two.sided"))
```

```
           Welch Two Sample t-test

data:  MUNICH$avg_temp and global$avg_temp
t = -57.705, df = 460.32, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.885643 -3.629709

sample estimates:
mean of x mean of y
4.611798  8.369474
```