Empirical Project 1: Working in R code

# Empirical Project 1 Working in R

These code downloads have been constructed as supplements to the full Doing Economics projects (<https://core-econ.org/doing-economics/>). You’ll need to download the data before running the code that follows.

## Part 1.1 The behaviour of average surface temperature over time

#### 1. In this dataset, temperature is measured as ‘anomalies’ rather than as absolute temperature. Using NASA’s Frequently Asked Questions section as a reference, explain in your own words what temperature ‘anomalies’ means. Why have researchers chosen this particular measure over other measures (such as absolute temperature)?

Las anomalías en la temperatura indican la diferencia en la temperatura de un período específico respecto a un valor fijo: el promedio desde 19551 a 1980. Se mide en anomalías y no en temperatura absoluta porque esta varía mucho en distancias cortas, mientras que las anomalías representan una región amplia.

### R walk-through 1.1 Importing the datafile into R

We want to import the datafile called ‘NH.Ts+dSST.csv’ into R.

We start by setting our working directory using the setwd command. This command tells R where your datafiles are stored. In the code below, replace ‘YOURFILEPATH’ with the full filepath that indicates the folder in which you have saved the datafile. If you don’t know how to find the path to your working folder, see the ‘Technical Reference’ section (<https://tinyco.re/3407438>).

Since our data is in csv format, we use the read.csv function to import the data into R. We will call our file ‘tempdata’ (short for ‘temperature data’).

Here you can see commands to R which are spread across two lines. You can spread a command across multiple lines, but you must adhere to the following two rules for this to work. First, the line break should come inside a set of parenthesis (i.e. between ( and ) or straight after the assignment operator (<-). Second, the line break must not be inside a string (whatever is inside quotes) or in the middle of a word or number.

tempdata <- "https://raw.githubusercontent.com/isismrtz/Taller-3/refs/heads/main/Datos%20Originales/NH.Ts%2BdSST.csv"  
  
tempdata <- read.csv(tempdata,  
 skip = 1, na.strings = "\*\*\*")

When using this function, we added two options. If you open the spreadsheet in Excel, you will see that the real data table only starts in Row 2, so we use the skip = 1 option to skip the first row when importing the data. When looking at the spreadsheet, you can see that missing temperature data is coded as "\*\*\*". In order for R to recognise the non-missing temperature data as numbers, we use the na.strings = "\*\*\*" option to indicate that missing observations in the spreadsheet are coded as "\*\*\*".

To check that the data has been imported correctly, you can use the head function to view the first six rows of the dataset, and confirm that they correspond to the columns in the csv file.

# Escribe aqui el codigo

Before working with the important data, we use the str function to check that the data is formatted correctly.

str(tempdata)

## 'data.frame': 145 obs. of 19 variables:  
## $ Year: int 1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 ...  
## $ Jan : num -0.38 -0.3 0.26 -0.57 -0.15 -1 -0.73 -1.08 -0.48 -0.27 ...  
## $ Feb : num -0.52 -0.24 0.22 -0.65 -0.1 -0.45 -0.83 -0.7 -0.61 0.3 ...  
## $ Mar : num -0.23 -0.05 0.03 -0.14 -0.63 -0.23 -0.7 -0.44 -0.63 -0.01 ...  
## $ Apr : num -0.3 -0.02 -0.29 -0.29 -0.58 -0.48 -0.36 -0.38 -0.21 0.17 ...  
## $ May : num -0.04 0.05 -0.22 -0.25 -0.35 -0.57 -0.33 -0.25 -0.14 -0.03 ...  
## $ Jun : num -0.18 -0.33 -0.28 -0.11 -0.4 -0.44 -0.37 -0.2 -0.02 -0.06 ...  
## $ Jul : num -0.21 0.1 -0.28 -0.05 -0.4 -0.33 -0.14 -0.24 0 -0.08 ...  
## $ Aug : num -0.25 -0.05 -0.14 -0.22 -0.5 -0.4 -0.42 -0.55 -0.21 -0.2 ...  
## $ Sep : num -0.24 -0.27 -0.24 -0.33 -0.44 -0.38 -0.32 -0.2 -0.19 -0.29 ...  
## $ Oct : num -0.29 -0.44 -0.51 -0.15 -0.43 -0.36 -0.31 -0.49 -0.03 -0.41 ...  
## $ Nov : num -0.43 -0.36 -0.33 -0.43 -0.57 -0.37 -0.4 -0.27 0 -0.61 ...  
## $ Dec : num -0.41 -0.23 -0.67 -0.14 -0.46 -0.1 -0.21 -0.42 -0.23 -0.54 ...  
## $ J.D : num -0.29 -0.18 -0.2 -0.28 -0.42 -0.43 -0.43 -0.43 -0.23 -0.17 ...  
## $ D.N : num NA -0.19 -0.17 -0.32 -0.39 -0.46 -0.42 -0.42 -0.25 -0.14 ...  
## $ DJF : num NA -0.32 0.08 -0.63 -0.13 -0.64 -0.55 -0.66 -0.5 -0.07 ...  
## $ MAM : num -0.19 -0.01 -0.16 -0.23 -0.52 -0.42 -0.46 -0.35 -0.33 0.04 ...  
## $ JJA : num -0.21 -0.09 -0.23 -0.12 -0.44 -0.39 -0.31 -0.33 -0.07 -0.11 ...  
## $ SON : num -0.32 -0.36 -0.36 -0.31 -0.48 -0.37 -0.34 -0.32 -0.07 -0.44 ...

You can see that all variables are formatted as numerical data (num), so R correctly recognises that the data are numbers.

[End of walk-through]

### R walk-through 1.2 Drawing a line chart of temperature and time

The data is formatted as numerical (num) data, so R recognises each variable as a series of numbers (instead of text), but does not recognise that these numbers correspond to the same variable for different time periods (known as ‘time series data’ in economics). Letting R know that we have time series data will make coding easier later (especially with making graphs). You can use the ts function to specify that a variable is a time series. Make sure to amend the code below so that the end year (end = c()) corresponds to the latest year in your dataset (our example uses 2017).

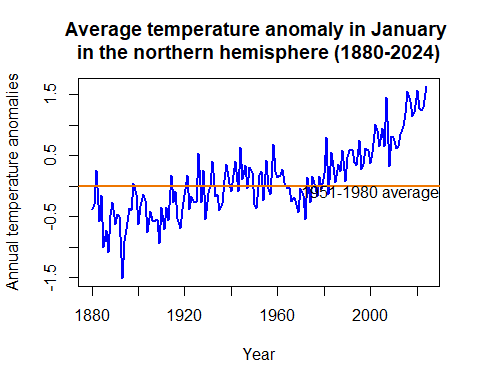
tempdata$Jan <- ts(tempdata$Jan,   
 start = c(1880), end = c(2024), frequency = 1)   
tempdata$DJF <- ts(tempdata$DJF,   
 start = c(1880), end = c(2024), frequency = 1)  
tempdata$MAM <- ts(tempdata$MAM,   
 start = c(1880), end = c(2024), frequency = 1)  
tempdata$JJA <- ts(tempdata$JJA,   
 start = c(1880), end = c(2024), frequency = 1)  
tempdata$SON <- ts(tempdata$SON,   
 start = c(1880), end = c(2024), frequency = 1)  
tempdata$J.D <- ts(tempdata$J.D,   
 start = c(1880), end = c(2024), frequency = 1)

Note that we placed each of these quarterly series in the relevant middle month. You could do the same for the remaining series, but we will only use the series above in this R walk-through.

We can now use these variables to draw line charts using the plot function. As an example, we will draw a line chart using data for January (tempdata$Jan) for the years 1880–2016. The title option on the next line adds a chart title, and the abline option draws a horizontal line according to our specifications. Make sure to amend the code below so that your chart title corresponds to the latest year in your dataset (our example uses 2016).

#### 2. Choose one month and plot a line chart with average temperature anomaly on the vertical axis and time (from 1880 to the latest year available) on the horizontal axis. Label each axis appropriately and give your chart a suitable title (refer to Figure 1.1 as an example.)

# Set line width and colour  
plot(tempdata$Jan, type = "l", col = "blue", lwd = 2,  
 ylab = "Annual temperature anomalies", xlab = "Year")  
  
# Add a title  
title("Average temperature anomaly in January \nin the northern hemisphere (1880-2024)")  
  
# Add a horizontal line (at y = 0)  
abline(h = 0, col = "darkorange2", lwd = 2)  
  
# Add a label to the horizontal line  
text(2000, -0.1, "1951-1980 average")

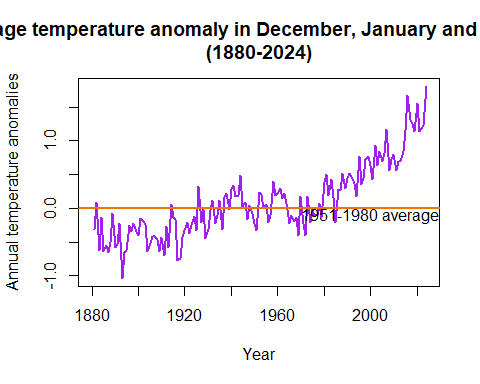


Try different values for type and col in the plot function to figure out what these options do (some online research could help). xlab and ylab define the respective axis titles.

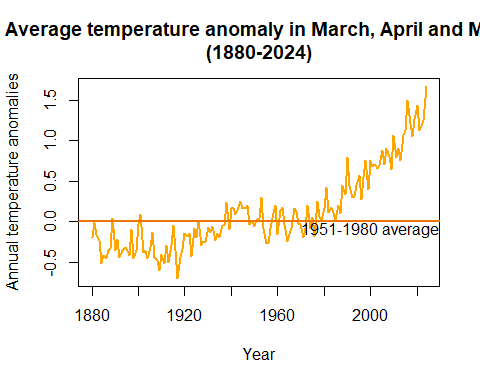
It is important to remember that all axis and chart titles should be enclosed in quotation marks (""), as well as any words that are not options (for example, colour names or filenames).

#### 3.Extra practice: The columns labelled DJF, MAM, JJA, and SON contain seasonal averages (means). For example, the MAM column contains the average of the March, April, and May columns for each year.Plot a separate line chart for each season, using average temperature anomaly for that season on the vertical axis and time (from 1880 to the latest year available) on the horizontal axis.

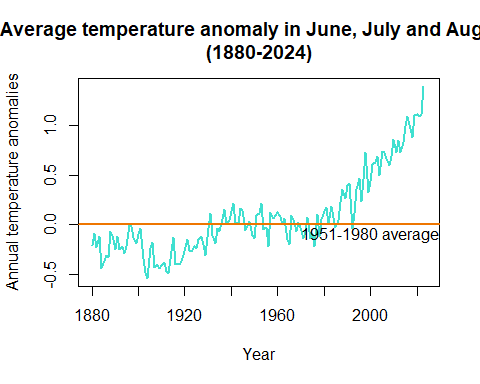
plot(tempdata$DJF, type = "l", col = "purple", lwd = 2,  
 ylab = "Annual temperature anomalies", xlab = "Year")  
  
# Add a title  
title("Average temperature anomaly in December, January and February\n(1880-2024)")  
abline(h = 0, col = "darkorange2", lwd = 2)  
  
# Add a label to the horizontal line  
text(2000, -0.1, "1951-1980 average")



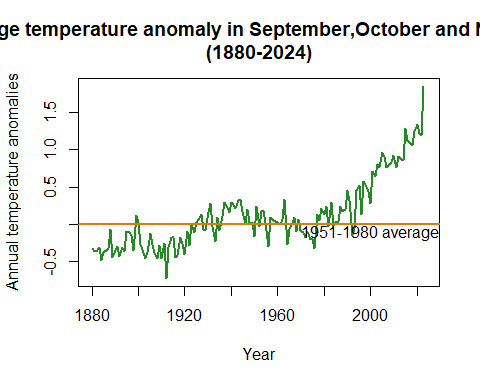
plot(tempdata$MAM, type = "l", col = "orange", lwd = 2,  
 ylab = "Annual temperature anomalies", xlab = "Year")  
  
# Add a title  
title("Average temperature anomaly in March, April and May \n(1880-2024)")  
abline(h = 0, col = "darkorange2", lwd = 2)  
# Add a label to the horizontal line  
text(2000, -0.1, "1951-1980 average")



plot(tempdata$JJA, type = "l", col = "turquoise", lwd = 2,  
 ylab = "Annual temperature anomalies", xlab = "Year")  
  
# Add a title  
title("Average temperature anomaly in June, July and August \n(1880-2024)")  
  
abline(h = 0, col = "darkorange2", lwd = 2)  
# Add a label to the horizontal line  
text(2000, -0.1, "1951-1980 average")



plot(tempdata$SON, type = "l", col = "forestgreen", lwd = 2,  
 ylab = "Annual temperature anomalies", xlab = "Year")  
  
# Add a title  
title("Average temperature anomaly in September,October and November \n(1880-2024)")  
  
abline(h = 0, col = "darkorange2", lwd = 2)  
text(2000, -0.1, "1951-1980 average")

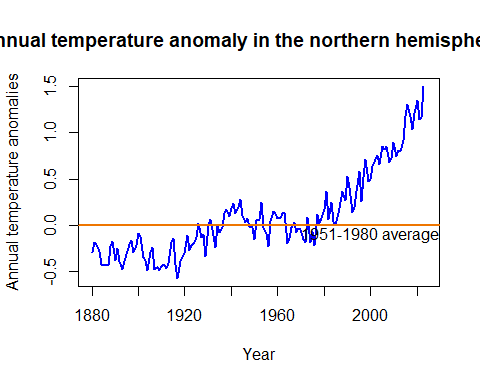


### R walk-through 1.3 Producing a line chart for the annual temperature anomalies

#### 4. The column labelled J–D contains the average temperature anomaly for each year.

##### a. Plot a line chart with annual average temperature anomaly on the vertical axis and time (from 1880 to the latest year available) on the horizontal axis. Your chart should look like Figure 1.1. Extension: Add a horizontal line that intersects the vertical axis at 0, and label it ‘1951–1980 average’.

# Set line width and colour  
plot(tempdata$J.D, type = "l", col = "blue", lwd = 2,  
 ylab = "Annual temperature anomalies", xlab = "Year")  
  
# Add a title  
title("Average annual temperature anomaly in the northern hemisphere (1880-2024)")  
  
# Add a horizontal line (at y = 0)  
abline(h = 0, col = "darkorange2", lwd = 2)  
  
# Add a label to the horizontal line  
text(2000, -0.1, "1951-1980 average")



##### b. What do your charts from Questions 2 to 4(a) suggest about the relationship between temperature and time?

Todos los graficos desde la pregunta 2 hasta la 4a), sugieren que la temperatura ha aumentado de una manera significativa con el paso de los años, especialmente despues de el año 2000, cuando se comienza a presentar una tendencia al alza sostenida e la temperatura, sin importar el trimestre que se esté graficando.

#### 5. You now have charts for three different time intervals: month (Question 2), season (Question 3), and year (Question 4). For each time interval, discuss what we can learn about patterns in temperature over time that we might not be able to learn from the charts of other time intervals.

Periodo mensual:

-Permite identificar anomalías de temperatura en períodos cortos de tiempo.

-Es Útil para analizar fenómenos climáticos específicos, como olas de calor o eventos de El Niño/La Niña.

-Puede mostrar fluctuaciones naturales que no afectan la tendencia a largo plazo, lo que podría dificultar la identificación del calentamiento global sin un análisis adicional.

Periodo Estacional (Jun-Ago):

-Permite observar tendencias específicas en los meses de verano (del hemisferio norte).

-Puede ayudar a identificar si las estaciones cálidas se están volviendo más extremas en comparación con otras épocas del año.

-Está un poco sesgada, pues no muestra el panorama completo del calentamiento global durante todo el año, pues solo cuenta con la infoermacion de 3 meses. Lo mismo ocurre para los otros grupos de trimestres del año.

Periodo anual:

-Muestra la tendencia general del cambio climático a largo plazo desde 1880 hasta 2024.

-Permite ver el cambio del calentamiento global de manera más clara y real, sin el ruido de variaciones estacionales o mensuales.

-Se puede observar la tendencia ascendente en las anomalías de temperatura con mayor claridad.

-No captura la variabilidad dentro de cada año, lo que podría ocultar eventos extremos de temperatura dentro de un periodo pequeño de tiempo.

#### 6. Compare your chart from Question 4 to Figure 1.4, which also shows the behaviour of temperature over time using data taken from the National Academy of Sciences.

##### a. Discuss the similarities and differences between the charts. (For example, are the horizontal and vertical axes variables the same, or do the lines have the same shape?)

Ambos gráficos muestran la anomalía de temperatura a lo largo del tiempo. En el eje horizontal, el primer gráfico abarca los años 1000-2000, mientras que el segundo cubre 1880-2020

Mientras las anteriores muestran calentamiento desde 1880, esta confirma que el aumento actual no tiene precedentes en los últimos 1000 años. Las variaciones pasadas fueron graduales y reversibles, pero el calentamiento reciente es rápido y extremo.

En el eje vertical, ambos gráficos representan desviaciones de temperatura respecto a un período base (1961-1990 en el primer gráfico y 1951-1980 en el segundo).

Otras consideraciones adicionales que se pueden tener en cuenta son:

Óptimo Climático Medieval (1000-1300): Período cálido moderado. Pequeña Edad de Hielo (1300-1800): Período prolongado de enfriamiento. Siglo XIX-XX: Recuperación gradual hasta el aumento extremo reciente.

##### b. Looking at the behaviour of temperature over time from 1000 to 1900 in Figure 1.4, are the observed patterns in your chart unusual?

En el primer gráfico, la temperatura muestra variaciones moderadas entre 1000 y 1900, con oscilaciones en periodos específicos, pero un aumento drástico después de 1900. En el segundo gráfico, la tendencia de la temperatura después de 1880 es similar a la parte final del primer gráfico, con un aumento evidente a partir de la segunda mitad del siglo XX.

##### c. Based on your answers to Questions 4 and 5, do you think the government should be concerned about climate change?

El gobierno si debe preocuparse por el cambio climatico pues los gráficos anteriores muestran que el calentamiento reciente es mucho más abrupto que las variaciones históricas. Esto sugiere que factores como la industria han acrecentado la emisión de gases de efecto invernadero, los cuales están contribuyendo significativamente al cambio climático. Dado el impacto potencial en los ecosistemas, la economía y la calidad de vida de la población, es crucial que los gobiernos implementen políticas para mitigar y adaptarse al calentamiento global.

[End of walk-through]

## Part 1.2 Variation in temperature over time

This is where the power of programming languages becomes evident: to produce the same line chart for a different variable, we simply take the code used in R walk-through 1.2 and replace the variable name Jan with the name for the annual variable (J.D). Again, make sure to amend the code so that your chart title corresponds to the latest year in your data (our example uses 2016).

### R walk-through 1.4 Creating frequency tables and histograms

Since we will be looking at data from different subperiods (year intervals) separately, we will create a categorical variable (a variable that has two or more categories) that indicates the subperiod for each observation (row). In R this type of variable is called a ‘factor variable’. When we create a factor variable, we need to define the categories that this variable can take.

tempdata$Period <-   
 factor(NA, levels =   
 c("1921-1950", "1951-1980", "1981-2010"),   
 ordered = TRUE)

We created a new variable called Period and defined the possible categories (which R refers to as ‘levels’). Since we will not be using data for some years (before 1921 and after 2010), we want Period to take the value ‘NA’ (not available) for these observations (rows), and the appropriate category for all the other observations (between 1921–2010). One way to do this is by defining Period as ‘NA’ for all observations, then change the values of Period for the observations in 1921–2010.

tempdata$Period[(tempdata$Year > 1920) &  
 (tempdata$Year < 1951)] <- "1921-1950"  
tempdata$Period[(tempdata$Year > 1950) &  
 (tempdata$Year < 1981)] <- "1951-1980"  
tempdata$Period[(tempdata$Year > 1980) &  
 (tempdata$Year < 2011)] <- "1981-2010"

We need to use all monthly anomalies from June, July, and August, but they are currently in three separate columns. We will use the c (combine) function to create one new variable (called temp\_summer) that contains all these values.

# Combine the temperature data for June, July, and August  
temp\_summer <- c(tempdata$Jun, tempdata$Jul, tempdata$Aug)

Now we have one long variable (temp\_summer), with the monthly temperature anomalies for the three months (from 1880 to the latest year) attached to each other. But remember that we want to make separate calculations for each category in Period (1921–1950, 1951–1980, 1981–2010). To make a variable showing the categories for the temp\_summer variable, we use the c function again.

temp\_summer <- unlist(tempdata[,7:9],use.names = FALSE)

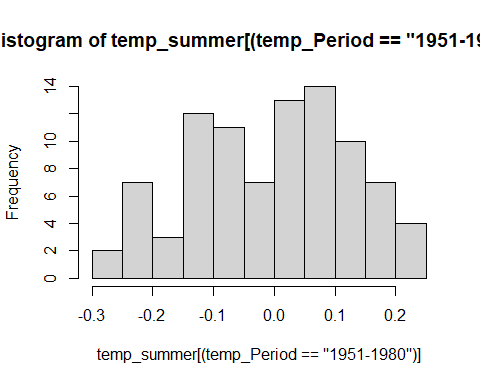
# Mirror the Period information for temp\_sum  
temp\_Period <-   
c(tempdata$Period, tempdata$Period, tempdata$Period)  
  
# Repopulate the factor information   
#temp\_Period <- factor(temp\_Period,   
# levels = 1:nlevels(tempdata$Period),   
# labels = levels(tempdata$Period))

After using the c function, we had to use the factor function again to tell R that our new variable temp\_Period is a factor variable.

We have now created the variables needed to make frequency tables and histograms (temp\_summer and temp\_Period). To obtain the frequency table for 1951–1980, we use the hist function on the monthly temperature anomalies from the period ‘1951–1980’: temp\_summer[(temp\_Period == "1951-1980")]. The option plot = FALSE tells R not to make a plot of this information. (See what happens if you set it to TRUE.)

#### 1. Using the monthly data for June, July, and August, create two frequency tables similar to Figure 1.5 for the years 1951–1980 and 1981–2010 respectively. The values in the first column should range from −0.3 to 1.05, in intervals of 0.05. See R walk-through 1.4 for how to do this.

hist(temp\_summer[(temp\_Period == "1951-1980")],plot=TRUE)



# Definir los rangos de temperatura  
breaks <- seq(-0.3, 1.05, by = 0.05)  
  
# Filtrar datos por periodo  
temp\_1951\_1980 <- temp\_summer[temp\_Period == "1951-1980"]  
temp\_1981\_2010 <- temp\_summer[temp\_Period == "1981-2010"]  
  
# Aplicar cut() antes de usar table()  
freq\_1951\_1980 <- table(cut(temp\_1951\_1980, breaks = breaks, include.lowest = TRUE))  
freq\_1981\_2010 <- table(cut(temp\_1981\_2010, breaks = breaks, include.lowest = TRUE))  
  
# Convertir en dataframes  
freq\_table\_1951\_1980 <- data.frame(Range = names(freq\_1951\_1980), Frequency = as.numeric(freq\_1951\_1980))  
freq\_table\_1981\_2010 <- data.frame(Range = names(freq\_1981\_2010), Frequency = as.numeric(freq\_1981\_2010))  
  
# Mostrar resultados  
print(freq\_table\_1951\_1980)

## Range Frequency  
## 1 [-0.3,-0.25] 2  
## 2 (-0.25,-0.2] 7  
## 3 (-0.2,-0.15] 3  
## 4 (-0.15,-0.1] 12  
## 5 (-0.1,-0.05] 11  
## 6 (-0.05,5.55e-17] 7  
## 7 (5.55e-17,0.05] 13  
## 8 (0.05,0.1] 14  
## 9 (0.1,0.15] 10  
## 10 (0.15,0.2] 7  
## 11 (0.2,0.25] 4  
## 12 (0.25,0.3] 0  
## 13 (0.3,0.35] 0  
## 14 (0.35,0.4] 0  
## 15 (0.4,0.45] 0  
## 16 (0.45,0.5] 0  
## 17 (0.5,0.55] 0  
## 18 (0.55,0.6] 0  
## 19 (0.6,0.65] 0  
## 20 (0.65,0.7] 0  
## 21 (0.7,0.75] 0  
## 22 (0.75,0.8] 0  
## 23 (0.8,0.85] 0  
## 24 (0.85,0.9] 0  
## 25 (0.9,0.95] 0  
## 26 (0.95,1] 0  
## 27 (1,1.05] 0

print(freq\_table\_1981\_2010)

## Range Frequency  
## 1 [-0.3,-0.25] 0  
## 2 (-0.25,-0.2] 0  
## 3 (-0.2,-0.15] 0  
## 4 (-0.15,-0.1] 1  
## 5 (-0.1,-0.05] 3  
## 6 (-0.05,5.55e-17] 2  
## 7 (5.55e-17,0.05] 6  
## 8 (0.05,0.1] 4  
## 9 (0.1,0.15] 3  
## 10 (0.15,0.2] 4  
## 11 (0.2,0.25] 4  
## 12 (0.25,0.3] 7  
## 13 (0.3,0.35] 7  
## 14 (0.35,0.4] 2  
## 15 (0.4,0.45] 7  
## 16 (0.45,0.5] 8  
## 17 (0.5,0.55] 5  
## 18 (0.55,0.6] 2  
## 19 (0.6,0.65] 4  
## 20 (0.65,0.7] 9  
## 21 (0.7,0.75] 6  
## 22 (0.75,0.8] 1  
## 23 (0.8,0.85] 2  
## 24 (0.85,0.9] 3  
## 25 (0.9,0.95] 0  
## 26 (0.95,1] 0  
## 27 (1,1.05] 0

From the output you can see that we can get the temperature ranges (the values in $breaks correspond to Column 1 of Figure 1.5) and the frequencies ($counts), which is all we need to create a frequency table. However, in our case the frequency table is merely a temporary input required to produce a histogram.

We can make the three histograms we need all at once, using the histogram function from the mosaic package.

The function below includes multiple commands:

* | temp\_Period splits the data according to its category, given by temp\_Period.
* type = "count" indicates that we want to display the counts (frequencies) in each category.
* breaks = seq(-0.5, 1.3, 0.1) gives a sequence of numbers −0.5, −0.4, …, 1.3, which are boundaries for the categories.
* main = "Histogram of temperature anomalies" gives Figure 1.6 its title.

#### 2.Using the frequency tables from Question 1:

##### a) Plot two separate column charts (frequency histograms) for 1951–1980 and 1981–2010 to show the distribution of temperatures, with frequency on the vertical axis and the range of temperature anomaly on the horizontal axis.Your charts should look similar to those in the New York Times article.

# Load the library we use for the following command.  
library(mosaic)

## Warning: package 'mosaic' was built under R version 4.3.3

## Registered S3 method overwritten by 'mosaic':  
## method from   
## fortify.SpatialPolygonsDataFrame ggplot2

##   
## The 'mosaic' package masks several functions from core packages in order to add   
## additional features. The original behavior of these functions should not be affected by this.

##   
## Attaching package: 'mosaic'

## The following objects are masked from 'package:dplyr':  
##   
## count, do, tally

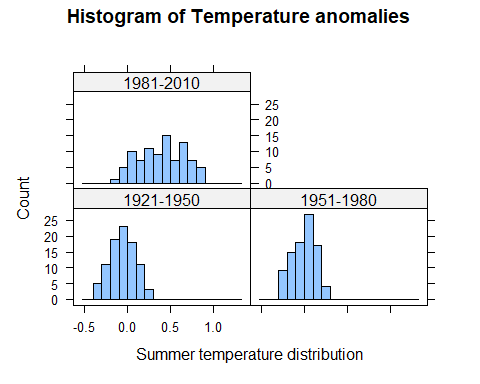
## The following object is masked from 'package:Matrix':  
##   
## mean

## The following object is masked from 'package:ggplot2':  
##   
## stat

## The following objects are masked from 'package:stats':  
##   
## binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,  
## quantile, sd, t.test, var

## The following objects are masked from 'package:base':  
##   
## max, mean, min, prod, range, sample, sum

histogram(~ temp\_summer | temp\_Period, type = "count",   
 breaks = seq(-0.5, 1.3, 0.10),   
 main = "Histogram of Temperature anomalies",   
 xlab = "Summer temperature distribution")



##### b) Using your charts, describe the similarities and differences (if any) between the distributions of temperature anomalies in 1951–1980 and 1981–2010.

Las distribuciones de anomalías de temperatura para los períodos 1951-1980 y 1981-2010 presentan diferencias notables:

Desplazamiento hacia valores más altos: En 1951-1980, la distribución de anomalías de temperatura está centrada alrededor de 0, con la mayoría de los valores en un rango cercano a -0.5 y 0.5. En contraste, en 1981-2010, la distribución se ha desplazado hacia valores positivos, con la mayor parte de los datos en el rango de 0.0 a 1.0.

Cambio en la forma de la distribución: Mientras que la distribución en 1951-1980 es más simétrica, la de 1981-2010 muestra una mayor frecuencia de valores positivos, indicando un aumento en las anomalías de temperatura.

Reducción de valores negativos: En el período 1951-1980 aún se observan algunas anomalías de temperatura negativas, aunque en menor proporción. En 1981-2010, los valores negativos prácticamente desaparecen, reflejando un aumento generalizado de las temperaturas.

En términos de similitud, ambas distribuciones mantienen una forma unimodal y siguen un patrón de distribución relativamente concentrado alrededor de sus respectivas medias. Sin embargo, el desplazamiento hacia valores más altos en 1981-2010 indica un calentamiento significativo en comparación con el período anterior.

### R walk-through 1.5 Using the quantile function

First, we need to create a variable that contains all monthly anomalies in the years 1951–1980. Then, we use R’s quantile function to find the required percentiles (0.3 and 0.7 refer to the 3rd and 7th deciles, respectively).

*Note*: You may get slightly different values to those shown here if you are using the latest data.

# Select years 1951 to 1980  
temp\_all\_months <- subset(tempdata,   
 (Year >= 1951 & Year <= 1980))  
   
# Columns 2 to 13 contain months Jan to Dec.  
temp\_51to80 <- unlist(temp\_all\_months[, 2:13])  
   
# c(0.3, 0.7) indicates the chosen percentiles.  
perc <- quantile(temp\_51to80, c(0.3, 0.7))   
  
# The cold threshold  
p30 <- perc[1]  
p30

## 30%   
## -0.1

# The hot threshold  
p70 <- perc[2]  
p70

## 70%   
## 0.1

#### 3.The New York Times article considers the bottom third (the lowest or coldest one-third) of temperature anomalies in 1951–1980 as ‘cold’ and the top third (the highest or hottest one-third) of anomalies as ‘hot’. In decile terms, temperatures in the 1st to 3rd decile are ‘cold’ and temperatures in the 7th to 10th decile or above are ‘hot’ (rounded to the nearest decile). Use R’s quantile function to determine what values correspond to the 3rd and 7th decile across all months in 1951–1980.

La afirmación mencionada anteriormente, es comprobada a partir de los datos pues en el tercer decil el valor es de -0,1 es decir las temperaturas hasta ese decil son frias y el valor del septimo decil es 0,1 lo que quiere decir que las tremperaturas superiores son “calientes”.

#### 4.Based on the values you found in Question 3, count the number of anomalies that are considered ‘hot’ in 1981–2010, and express this as a percentage of all the temperature observations in that period. Does your answer suggest that we are experiencing hotter weather more frequently in 1981–2010? (Remember that each decile represents 10% of observations, so 30% of temperatures were considered ‘hot’ in 1951–1980.)

Al analizar los datos se encuentra que alrededor del 30% de las observaciones tienden a mayor o igual al 0.1, esto indica que el clima calido ha sido experimentado con mayor frecuencia en el periodo de 1981-2010 a comparacion con los años de 1951-1980

[End of walk-through]

### R walk-through 1.6 Using the mean function

*Note*: You may get slightly different values to those shown here if you are using the latest data.

We repeat the steps used in R walk-through 1.5, now looking at monthly anomalies in the years 1981–2010. We can simply change the year values in the code from R walk-through 1.5.

# Select years 1951 to 1980  
temp\_all\_months <- subset(tempdata,   
 (Year >= 1981 & Year <= 2010))  
   
# Columns 2 to 13 contain months Jan to Dec.  
temp\_81to10 <- unlist(temp\_all\_months[, 2:13])

Now that we have all the monthly data for 1981–2010, we want to count the proportion of observations that are smaller than –0.1. This is easily achieved with the following lines of code:

paste("Proportion smaller than p30")

## [1] "Proportion smaller than p30"

temp <- temp\_81to10 < p30  
mean(temp)

## [1] 0.01944444

Let’s check whether we get a similar result for the number of observations that are larger than 0.11.

paste("Proportion larger than p70")

## [1] "Proportion larger than p70"

temp <- temp\_81to10 > p70  
mean(temp)

## [1] 0.8472222

[End of walk-through]

### R walk-through 1.7 Calculating and understanding mean and variance

Calculate mean and variance. One option is to use the mosaic package

paste("Mean of DJF temperature anomalies across periods")

## [1] "Mean of DJF temperature anomalies across periods"

mean(~DJF|Period,data=tempdata)

## 1921-1950 1951-1980 1981-2010   
## -0.030333333 -0.002666667 0.523333333

paste("Variance of DJF anomalies across periods")

## [1] "Variance of DJF anomalies across periods"

var(~DJF|Period,data=tempdata)

## 1921-1950 1951-1980 1981-2010   
## 0.05672057 0.05038575 0.07871264

Using the data in tempdata (data = tempdata), we calculated the mean (mean) and variance (var) of variable ~DJF separately for (|) each value of Period. The mosaic package allows us to calculate the means/variances for each period all at once. If mosaic is not loaded, you will get the error message: Error in mean(~DJF \| Period, data = tempdata) : unused argument (data = tempdata).

Looking at the results, it appears that it is not only the mean (December, January, and February) temperature anomaly that increases through 1981–2010, but also the variance.

Let’s calculate the variances through the periods for the other seasons.

#### 5. The New York Times article discusses whether temperatures have become more variable over time. One way to measure temperature variability is by calculating the variance of the temperature distribution.For each season (DJF, MAM, JJA, and SON):

##### a) Calculate the mean (average) and variance separately for the following time periods: 1921–1950, 1951–1980, and 1981–2010.

paste("Variance of MAM anomalies across periods")

## [1] "Variance of MAM anomalies across periods"

var(~MAM|Period,data=tempdata)

## 1921-1950 1951-1980 1981-2010   
## 0.03099782 0.02540000 0.07573345

paste("Variance of JJA anomalies across periods")

## [1] "Variance of JJA anomalies across periods"

var(~JJA|Period,data=tempdata)

## 1921-1950 1951-1980 1981-2010   
## 0.02128920 0.01460644 0.06749609

paste("Variance of SON anomalies across periods")

## [1] "Variance of SON anomalies across periods"

var(~SON|Period,data=tempdata)

## 1921-1950 1951-1980 1981-2010   
## 0.02819264 0.02635126 0.11104644

##### b) For each season, compare the variances in different periods, and explain whether or not temperature appears to be more variable in later periods.

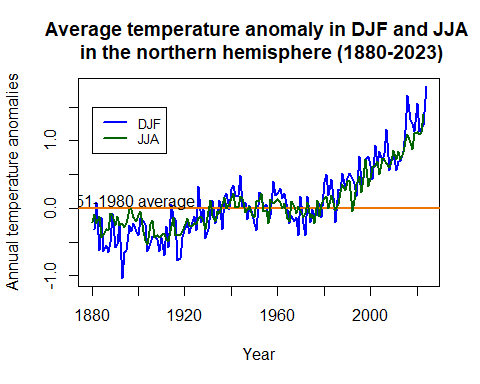
Si, la varianza ha aumentado en el período 1981-2010 en comparación con 1951-1980 y 1921-1950, esto respaldaría la hipótesis de que las temperaturas se han vuelto más variables con el tiempo, lo que coincide con la preocupación sobre el impacto del cambio climático en la estabilidad del clima.

#### 6. Using the findings of the New York Times article and your answers to Questions 1 to 5, discuss whether temperature appears to be more variable over time. Would you advise the government to spend more money on mitigating the effects of extreme weather events?

Dados los datos acerca del cambio climático esto se deduce en que el incremento en temperatura evidencia un fenomeno climatico extremo por lo tanto se recomienda al gobierno invertir en infraestructura para la proteccion de poblaciones vulnerables a inundaciones y politicas de reduccion de emision de efecto carbono.

We can plot a line chart to see these changes graphically. (This type of chart is formally known as a ‘time-series plot’). Make sure to change the chart title according to the latest year in your data (here we used 2016).

plot(tempdata$DJF, type = "l", col = "blue", lwd = 2,  
 ylab = "Annual temperature anomalies", xlab = "Year")  
  
# \n creates a line break  
title("Average temperature anomaly in DJF and JJA \n in the northern hemisphere (1880-2023)")  
  
# Add a horizontal line (at y = 0)  
abline(h = 0, col = "darkorange2", lwd = 2)  
lines(tempdata$JJA, col = "darkgreen", lwd = 2)   
  
# Add a label to the horizontal line  
text(1895, 0.1, "1951-1980 average")  
legend(1880, 1.5, legend = c("DJF", "JJA"),  
 col = c("blue", "darkgreen"),   
 lty = 1, cex = 0.8, lwd = 2)



[End of walk-through]

## Part 1.3 Carbon emissions and the environment

#### 1. The CO2 data was recorded from one observatory in Mauna Loa. Using an Earth System Research Laboratory article as a reference, explain whether or not you think this data is a reliable representation of the global atmosphere.

Los datos que nos proporcionan de Mauna Loa si se consideran como una representacion fiable de la atmosfera global, ya que esta nos muestra la mezcla de gases a gran escala y muestra las demas tendencias de medicion globales y por ello esta fuente es relevante para el estudio del cambio climatico

#### 2. The variables trend and interpolated are similar, but not identical. In your own words, explain the difference between these two measures of CO2 levels. Why might there be seasonal variation in CO2 levels?

La interpolated se utiliza para reflejar las fluctuaciones reales del CO2 en las variabilidad estacional mientras que Trendy se utiliza para analizar el aumento global del CO2 sin contar las variaciones estacionales.

### R walk-through 1.8 Scatterplots and the correlation coefficient

First we will use the read.csv function to import the CO2 datafile into R, and call it CO2data.

library(readxl)  
  
# Descargar el archivo desde GitHub  
url <- "https://github.com/isismrtz/Taller-3/blob/main/Datos%20Originales/doing-economics-datafile-working-in-excel-project-1.xlsx?raw=true"  
destfile <- "doing-economics-datafile-working-in-excel-project-1.xlsx"  
download.file(url, destfile, mode = "wb")  
  
# Leer el archivo Excel como un dataframe  
CO2data <- read\_excel(destfile)

This file has monthly data, but in contrast to the data in tempdata, the data is all in one column (this is more conventional than the column per month format). To make this task easier, we will pick the June data from the CO2 emissions and add them as an additional variable to the tempdata dataset.

R has a convenient function called merge to do this. First we create a new dataset that contains only the June emissions data (‘CO2data\_june’).

CO2data\_june <- CO2data[CO2data$Month == 6,]

Then we use this data in the merge function. The merge function takes the original ‘tempdata’ and the ‘CO2data’ and merges (combines) them together. As the two dataframes have a common variable, Year, R automatically matches the data by year.

(*Extension:* Look up ?merge or Google ‘How to use the R merge function’ to figure out what all.x does, and to see other options that this function allows.)

names(CO2data)[1] <- "Year"  
tempCO2data <- merge(tempdata,CO2data\_june)

Let us have a look at the data and check that it was combined correctly:

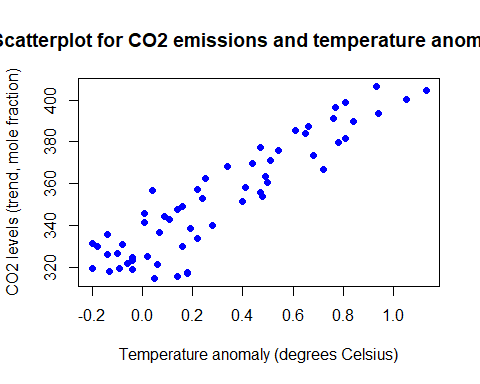
head(tempCO2data[, c("Year", "Jun", "Trend")])

## Year Jun Trend  
## 1 1958 0.05 314.85  
## 2 1959 0.14 315.92  
## 3 1960 0.18 317.36  
## 4 1961 0.18 317.48  
## 5 1962 -0.13 318.27  
## 6 1963 -0.04 319.16

## Year Jun Trend  
## 1 1958 0.04 314.85  
## 2 1959 0.14 315.92  
## 3 1960 0.18 317.36  
## 4 1961 0.19 317.48  
## 5 1962 -0.10 318.27  
## 6 1963 -0.02 319.16

To make a scatterplot, we use the plot function. R’s default chart for plot is a scatterplot, so we do not need to specify the chart type. One new option that applies to scatterplots is pch =, which determines the appearance of the data points. The number 16 corresponds to filled-in circles, but you can experiment with other numbers (from 0 to 25) to see what the data points look like.

plot(tempCO2data$Jun, tempCO2data$Trend,   
 xlab = "Temperature anomaly (degrees Celsius)",   
 ylab = "CO2 levels (trend, mole fraction)",   
 pch = 16, col = "blue",cex=1)  
  
title("Scatterplot for CO2 emissions and temperature anomalies")



The cor function calculates the correlation coefficient. *Note*: You may get slightly different results if you are using the latest data.

cor(tempCO2data$Jun,tempCO2data$Trend)

## [1] 0.9149093

cor(tempCO2data$Nov,tempCO2data$Trend)

## [1] 0.8575439

## [1] 0.9157744

In this case, the correlation coefficient tells us that the data is quite close to resembling an upward-sloping straight line (as seen on the scatterplot). There is a strong positive association between the two variables (higher temperature anomalies are associated with higher CO2 levels).

One limitation of this correlation measure is that it only tells us about the strength of the upward- or downward-sloping linear relationship between two variables, in other words how closely the scatterplot aligns along an upward- or downward-sloping straight line. The correlation coefficient cannot tell us if the two variables have a different kind of relationship (such as that represented by a wavy line).

*Note:* The word ‘strong’ is used for coefficients that are close to 1 or −1, and ‘weak’ is used for coefficients that are close to 0, though there is no precise range of values that are considered ‘strong’ or ‘weak’.

If you need more insight into correlation coefficients, you may find it helpful to watch online tutorials such as ‘Correlation coefficient intuition’ (<https://tinyco.re/4363520>) from the Khan Academy.

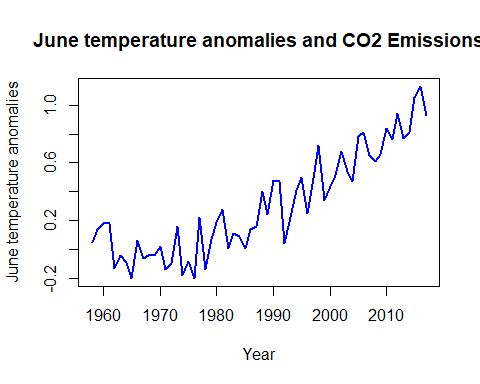
As we are dealing with time-series data, it is often more instructive to look at a line plot, as a scatterplot cannot convey how the observations relate to each other in the time dimension. If you were to check the variable types (using str(tempCO2data)), you would see that the data is not yet in time-series format. We could continue with the format as it is, but for plotting purposes it is useful to let R know that we are dealing with time-series data. We therefore apply the ts function as we did in Part 1.1.

tempCO2data$Jun <- ts(tempCO2data$Jun,   
 start = c(1958), end = c(2017), frequency = 1)   
tempCO2data$Trend <- ts(tempCO2data$Trend,   
 start = c(1958), end = c(2017), frequency = 1)

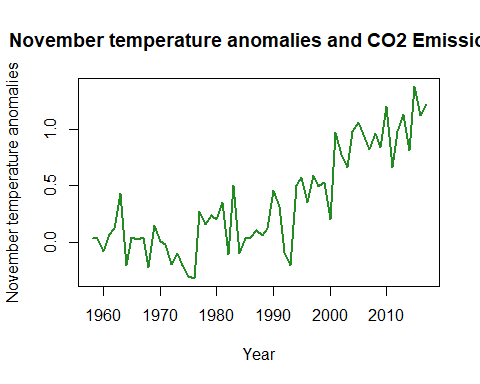
tempCO2data$Nov <- ts(tempCO2data$Nov,   
 start = c(1958), end = c(2017), frequency = 1)

Let’s start by plotting the June temperature anomalies.

plot(tempCO2data$Jun , col = "blue",lwd = 2,  
 ylab = "June temperature anomalies", xlab="Year")  
  
title ("June temperature anomalies and CO2 Emissions")

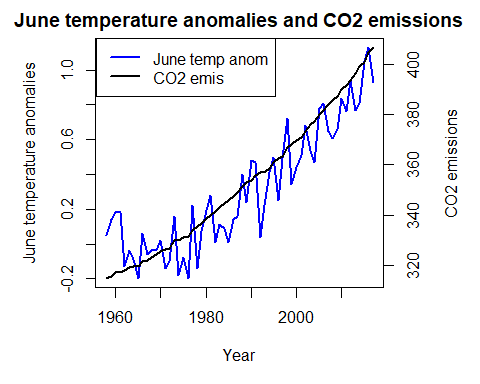


plot(tempCO2data$Nov , col = "forestgreen",lwd = 2,  
 ylab = "November temperature anomalies", xlab="Year")  
  
title ("November temperature anomalies and CO2 Emissions")

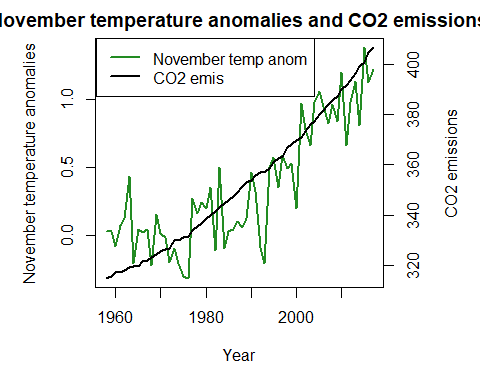


Typically, when using the plot function we would now only need to add the line for the second variable using the lines command. The issue, however, is that the CO2 emissions variable (Trend) is on a different scale, and the automatic vertical axis scale (from –0.2 to about 1.2) would not allow for the display of Trend. To resolve this issue you can introduce a second vertical axis using the commands below. (*Tip:* You are unlikely to remember the exact commands required, however you can Google ‘R plot 2 vertical axes’ or a similar search term, and then adjust the code you find so it will work on your dataset.)

# Create extra margins used for the second axis  
par(mar = c(5, 5, 2, 5))  
  
plot(tempCO2data$Jun, type = "l", col = "blue", lwd = 2,  
 ylab = "June temperature anomalies", xlab = "Year")  
  
title("June temperature anomalies and CO2 emissions")   
  
# This puts the next plot into the same picture.  
par(new = T)  
  
# No axis, no labels  
plot(tempCO2data$Trend, pch = 16, lwd = 2,   
 axes = FALSE, xlab = NA, ylab = NA, cex = 1.2)   
axis(side = 4)  
mtext(side = 4, line = 3, 'CO2 emissions')  
  
legend("topleft", legend = c("June temp anom", "CO2 emis"),  
 lty = c(1, 1), col = c("blue", "black"),lwd=2)



# Create extra margins used for the second axis  
par(mar = c(5, 5, 2, 5))  
  
plot(tempCO2data$Nov, type = "l", col = "forestgreen", lwd = 2,  
 ylab = "November temperature anomalies", xlab = "Year")  
  
title("November temperature anomalies and CO2 emissions")   
  
# This puts the next plot into the same picture.  
par(new = T)  
  
# No axis, no labels  
plot(tempCO2data$Trend, pch = 16, lwd = 2,   
 axes = FALSE, xlab = NA, ylab = NA, cex = 1.2)   
axis(side = 4)  
mtext(side = 4, line = 3, 'CO2 emissions')  
  
legend("topleft", legend = c("November temp anom", "CO2 emis"),  
 lty = c(1, 1), col = c("forestgreen", "black"),lwd=2)



This line graph not only shows how the two variables move together in general, but also clearly demonstrates that both variables display a clear upward trend over the sample period. This is an important feature of many (not all) time series variables, and is important for the interpretation (see the ‘Find out more’ box on spurious correlations that follows).

#### 3. Plot a line chart with interpolated and trend CO2 levels on the vertical axis and time (starting from January 1960) on the horizontal axis. Label the axes and the chart legend, and give your chart an appropriate title. What does this chart suggest about the relationship between CO2 and time?

El gráfico sugiere que los niveles de CO₂ han aumentado significativamente desde 1960 y que la tasa de crecimiento se mantiene constante o incluso acelerada. Esto indica una acumulación continua de CO₂ en la atmósfera.

#### 4. Choose one month and add the CO2 trend data to the temperature dataset from Part 1.1, making sure that the data corresponds to the correct year.

##### a) Make a scatterplot of CO2 level on the vertical axis and temperature anomaly on the horizontal axis.

##### b) Discuss the shortcomings of using this coefficient to summarize the relationship between variables.

El coeficiente de Pearson tiene varias limitaciones para resumir la relación entre variables:

-Relaciones no lineales: Solo captura relaciones lineales, por lo que no es útil para datos con asociaciones curvilíneas o complejas. -Sensibilidad a valores atípicos: Los valores extremos pueden distorsionar significativamente el coeficiente. -No implica causalidad: Aunque muestra una correlación, no indica si una variable causa la otra. -Requiere distribución normal: Asume que las variables siguen una distribución normal, lo que no siempre ocurre en la práctica.

#### 5. Extra practice: Choose two months and add the CO2 trend data to the temperature dataset from Part 1.1, making sure that the data corresponds to the correct year. Create a separate chart for each month. What do your charts and the correlation coefficients suggest about the relationship between CO2 levels and temperature anomalies?

Las gráficas muestran una tendencia creciente tanto en las anomalías de temperatura como en los niveles de CO2 a lo largo del tiempo, lo que visualmente refuerza la alta correlación observada en los coeficientes. La relación positiva fuerte (correlaciones cercanas a 1) sugiere que, a medida que el CO2 aumenta, las temperaturas también tienden a elevarse. Sin embargo, la correlación no implica causalidad, aunque la consistencia de las tendencias respalda una conexión significativa entre ambas variables

[End of walk-through]

#### 6. Consider the example of spurious correlation described above.

##### a) In your own words, explain spurious correlation and the difference between correlation and causation.

Una correlación espuria ocurre cuando dos variables parecen estar relacionadas, pero en realidad su relación es causada por un tercer factor no observado. La correlación indica que dos variables se mueven juntas, pero no implica que una cause la otra. La causalidad, en cambio, significa que un cambio en una variable provoca directamente un cambio en la otra.

##### b) Give an example of spurious correlation, similar to the one above, for either CO2 levels or temperature anomalies.

Podría existir una correlación entre el número de parques en una ciudad y el aumento de temperatura, pero esto no significa que los parques causen el calentamiento. Un tercer factor, como la urbanización, podría explicar la relación: más urbanización aumenta tanto la temperatura (por efecto isla de calor) como la construcción de parques para compensar la contaminación.

##### c) Choose an example of spurious correlation from Tyler Vigen’s website. Explain whether you think it is a coincidence, or whether this correlation could be due to one or more other variables.

En la correlación entre el número de películas en las que Orlando Bloom aparció y la calidad del aire en Union City es una coincidencia puesto que las variables no se relacionan entre si. No existiría ninguna variable que las pudiera conectar,