ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Desafío R (PAUTA)

Taller de análisis de datos I
Diplomado en Ciencia de Datos para Política Públicas

Fecha de entrega: Domingo 03/07/2022 hasta las 23:59 hrs.

Profesor: José Daniel Conejeros (jdconejeros@uc.cl)

Aspectos Formales (LEER)

1. Debe entregar una carpeta comprimida (`.zip` o `.rar`) con los siguientes archivos:

- El archivo del proyecto (`.Rproj`).
- Archivo de código `.R` con la función solicitada.
- Base de datos final en formato `.csv`. **No las bases por separado**.
- Las figuras generadas en formato `.png`.
- Documento `.pdf` con las figuras solicitadas y su interpretación.

No se aceptaran documentos en otros formatos.

2. Debe cargar este archivo a más tardar el Domingo 03/07/2022 a las 23:59 hrs en el buzón disponible en la plataforma del curso.

3. Este desafío se puede realizar en parejas (recomendable) o individual.

4. Dudas o solicitudes de reunión al correo electrónico: jdconejeros@uc.cl

Tendremos una sesión de dudas al final del taller 3 (25/06/2022)

No olvide identificar el documento con sus nombres y ruts

Desafío en R

El objetivo de este desafío es practicar funciones de `tidyverse` para la manipulación de tablas y uso de procesos iterativos con datos reales. Para esto usted deberá trabajar con la carpeta `data.zip` que contiene 284 tablas de datos (`.csv`) con el cambio de temperatura superficial promedio mensual por país entre 1961 y 2019. Cada archivo contiene las siguientes variables:

- `Area`: País o área geográfica.
- `Months`: Meses del año en inglés.
- `Y1961` hasta `Y2019`: Variables que indican el valor de la temperatura promedio observada en el años-mes.

Se requiere construir proceso automatizado en que lea el directorio donde se encuentran todos los datos, cargue las tablas y entregue el análisis solicitado. En concreto, se le pide que construya una función que reciba el nombre del país-area y un rango de tiempo en años (desde - hasta). Esta función debe realizar las siguientes tareas:

1. Cargar todas las tablas de datos que se encuentran en su directorio `.../data/`. *Hint:* Vea los ejemplos del taller 2 (3 puntos).

2. Una vez cargado los datos debe hacer lo siguiente de forma iterativa (`for`) sobre cada tabla:

a. Limpiar los nombres de las columnas para que queden en minúsculas. *Hint:* Revise las funciones `tolower()` o `str_to_lower()` (2 puntos).

b. Transformar las columnas `Y1961` hasta `Y2019` en filas, en otras palabras, se le pide transformar la tabla de formato wide a formato long. *Hint*: revise la función `pivot_longer()`. La variable con los años se debe llamar `year` y la variable con los valores de temperatura promedio se debe llamar `temperature_change` (10 puntos).

c. De la variable `year` generada en **b** debe sacar la expresión "Y" al principio de cada string. Por ejemplo, si el valor es `Y1961` usted se debe quedar con `1961`. *Hint:* revise las funciónes `mutate()`, `str_to_replace()` (5 puntos).

d. A partir de la variable `year` generada en **c** y `months`, usted debe construir una variable nueva llamada `date` con la siguiente estructura:

months	year	date
January	1961	1961-01-15

- La variable `date` debe ser de tipo "Date" puede verificar esto con la función `class()`). *Hint:* revise la funciones `mutate()`, `paste()` y `as.Date()` (5 puntos).

e. Una vez procesada cada tabla, debe unirlas por filas en un solo gran marco de datos con nombre `df` y guardarlo en su directorio en un archivo con extensión `.csv`. *Hint:* revise las funciones `bind_rows`, `add_rows`, etc. (5 puntos). La función debe retornar un marco de datos final con las siguientes columnas: `area`, `months`, `year`, `temperatura_change` y `date` (5 puntos).

3. Una vez generada su marco de datos en el ítem **2**. En la misma función usted debe construir un gráfico que muestre el cambio promedio de temperatura en un área y rango de tiempo determinado. La figura se debe guardar en su directorio de trabajo con la siguiente estructura: `area_desde_hasta.png`. *Hint*: Revise las funciones de `ggplot2` (15 puntos).

4. Aplique su función a los siguientes requerimientos de información (10 puntos):

- World: 1961 - 2019
- OECD: 1961 - 2019
- South America: 1961 - 2019
- Chile: 1961 - 2019

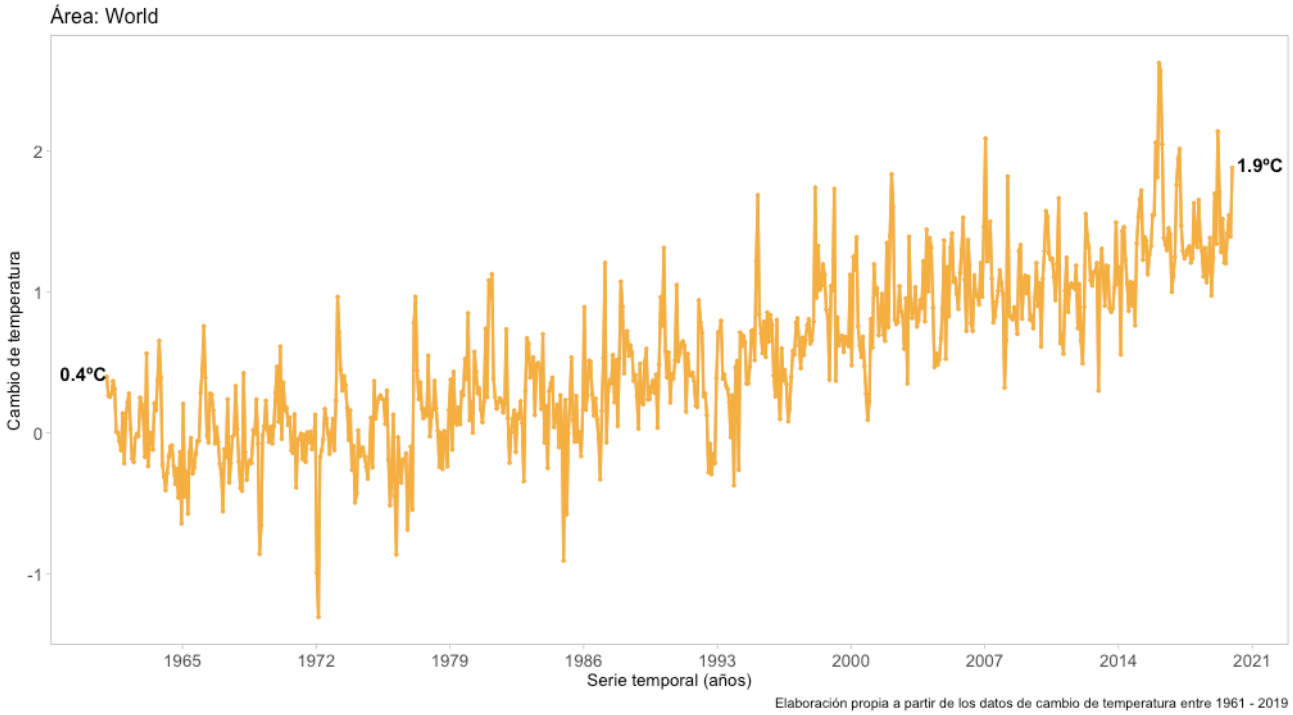
Guarde las figuras generadas en un pdf e interprete brevemente sus resultados. **¿Qué puede decir del cambio de temperatura en el tiempo para estas áreas?**

Resultado esperado

Si aplico mi función: `fun(areas="World", desde=1961, hasta=2019)` debería obtener mi base de datos (`df`) con toda la información:

```
  V1  area   months year temperature_change       date
1  1 World  January 1961              0.399 1961-01-15
2  2 World February 1961              0.263 1961-02-15
3  3 World    March 1961              0.254 1961-03-15
4  4 World    April 1961              0.278 1961-04-15
5  5 World      May 1961              0.367 1961-05-15
6  6 World     June 1961              0.313 1961-06-15
```

Y un gráfico similar a este (`.png`):



Código solución

Puede encontrar el código `Desafio_R_solucion.R` en: https://github.com/JDConejeros/DS_PP_UC/blob/main/APPD/Evaluaciones/Desafio_R_solucion.R

La función se espera que sea, más o menos, de la forma:

```
rm(list=(ls())) # Limpiamos el enviroment

fun <- function(areas, desde, hasta){

  # 0. Ajustes preliminares -----
  # Revisamos la info de la sesión
  sessionInfo()

  # Desactivar notación científica y limpiar la memoria
  options(scipen=999)

  Sys.setlocale(category = "LC_ALL", "en_US.UTF-8") # Ajuste de idioma

  # 0. Cargamos las librerías de trabajo -----
  # Vamos a instroducir una subfunción de instala/carga de librerías
  # Ajustes preliminares (paquetes)
  install_load <- function(packages){
    for (i in packages) {
      if (i %in% rownames(installed.packages())) {
        library(i, character.only=TRUE)
      } else {
        install.packages(i, repos = "http://cran.us.r-project.org")
        library(i, character.only = TRUE)
      }
    }
  }

  pkg <- c("rio", # Lectura de datos
           "dplyr", "tidyr", "stringr", # Manipulación de datos
           "ggplot2" # Visualización de datos
  )

  # Aplicamos nuestra subfunción
  install_load(pkg)

  # Lectura de la data

  # P1. Lectura de las bases de datos -----
  # Directorios complemetarios
  dir <- getwd()
  dir_files <- paste0(dir, "/APPD/Evaluaciones/data/")
  files <- list.files(dir_files, pattern = ".csv")

  df <- data.frame() # Data frame vacío auxiliar
```

```
data <- data.frame() # Data frame vacío con el resultado final

for(i in files){
  # Lectura de la data
  df <- import(paste0(dir_files, i))

  # P2. Procesamiento de datos -----

  ## P2a. -----
  colnames(df) <- tolower(colnames(df))
  colnames(df)

  ## P2b. -----
  df <- df %>%
    select(-v1) %>%
    pivot_longer(cols=!c("area", "months"), names_to="year", values_to = "temperature_change")

  ## P2c. -----
  df <- df %>%
    mutate(year=str_replace(year,"y",""))

  ## P2d. -----
  df <- df %>%
    mutate(date=as.Date(paste(months, "15", year), format="%b%d%Y"))

  ## P2e. -----
  df <- df %>%
    select(area, months, year, temperature_change, date)

  data <- rbind(data, df)
}

# Devolvemos la tabla final al enviroment
assign("df", data, 1)

# P3. Construímos nuestro gráfico -----

fig <- data %>%
  filter(area==areas & as.numeric(year)>=desde & as.numeric(year)<=hasta) %>%
  ggplot(aes(x=date, y=temperature_change, group=1)) +
  geom_point(aes(y=temperature_change), shape=1, size=0.5, color="#F5B041") +
  geom_line(aes(y=temperature_change), size=0.75, linetype=1, color="#F5B041") +
  geom_text(data = . %>% filter(date == max(date)),
            aes(label = paste0(round(temperature_change, 1), "ºC")),
            vjust = 0.3, hjust = -0.1,
            show.legend = FALSE, fontface="bold") +
  geom_text(data = . %>% filter(date == min(date)),
            aes(label = paste0(round(temperature_change, 1), "ºC")),
            vjust = 0.3, hjust = 1,
```

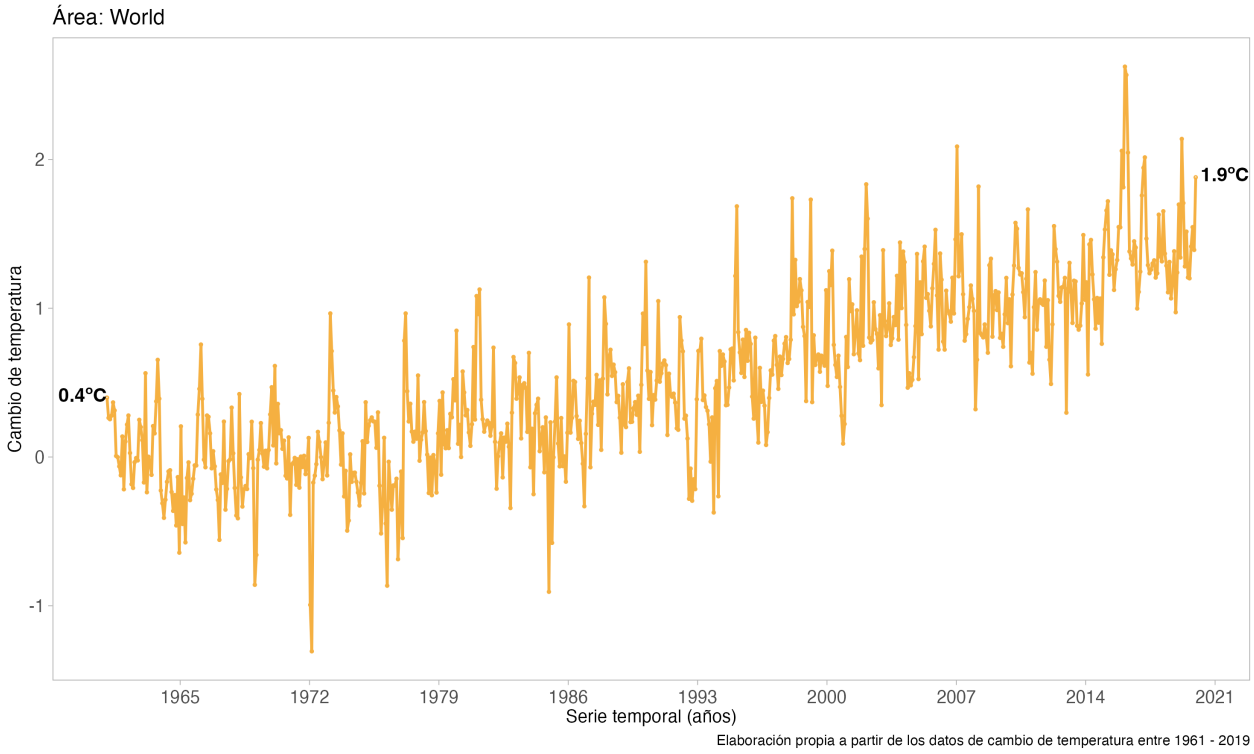
```
              show.legend = FALSE, fontface="bold") +
    scale_x_date(breaks = "7 year",
                 date_labels = "%Y") +
    labs(x="Serie temporal (años)",
         y = "Cambio de temperatura",
         title=paste("Área:", areas),
         caption="Elaboración propia a partir de los datos de cambio de temperatura entre 1961 - 2019") +
    theme_light(base_size = 10) +
    theme(axis.text.x=element_text(size=10,  vjust=0.5, angle = 0),
          axis.text.y=element_text(size=10),
          plot.title = element_text(size=12, hjust=0.0),
          legend.position="none",
          strip.text = element_text(size = 12, face = "bold", color = "gray40"),
          strip.background = element_rect(fill="white", colour="gray60", linetype="solid"),
          panel.grid.minor = element_blank(),
          panel.grid.major = element_blank())

  fig

  # Guardamos la figura
  # Debes crear un subdirectorio llamado evaluaciones y otro que se llame figuras_solucion
  subdir <-  "/APPD/Evaluaciones/resultados/grafico_"

  ggsave(plot = fig,
         filename = paste0(dir, subdir, areas, ".png"),
         res = 300,
         width = 25,
         height = 15,
         units = 'cm',
         scaling = 1,
         device = ragg::agg_png)

  # Guardamos la data
  subdir <-  "/APPD/Evaluaciones/resultados/"
  write.csv(data, paste0(dir, subdir, "df", ".csv"))


  # Removemos los objetos secundarios
  rm(dir, dir_files, files, i)

  # FIN DE LA FUNCIÓN -----
}
```

Gráficos

Puede encontrar los insumos resultados en: https://github.com/JDConejeros/DS_PP_UC/tree/main/APPD/Evaluaciones/resultados

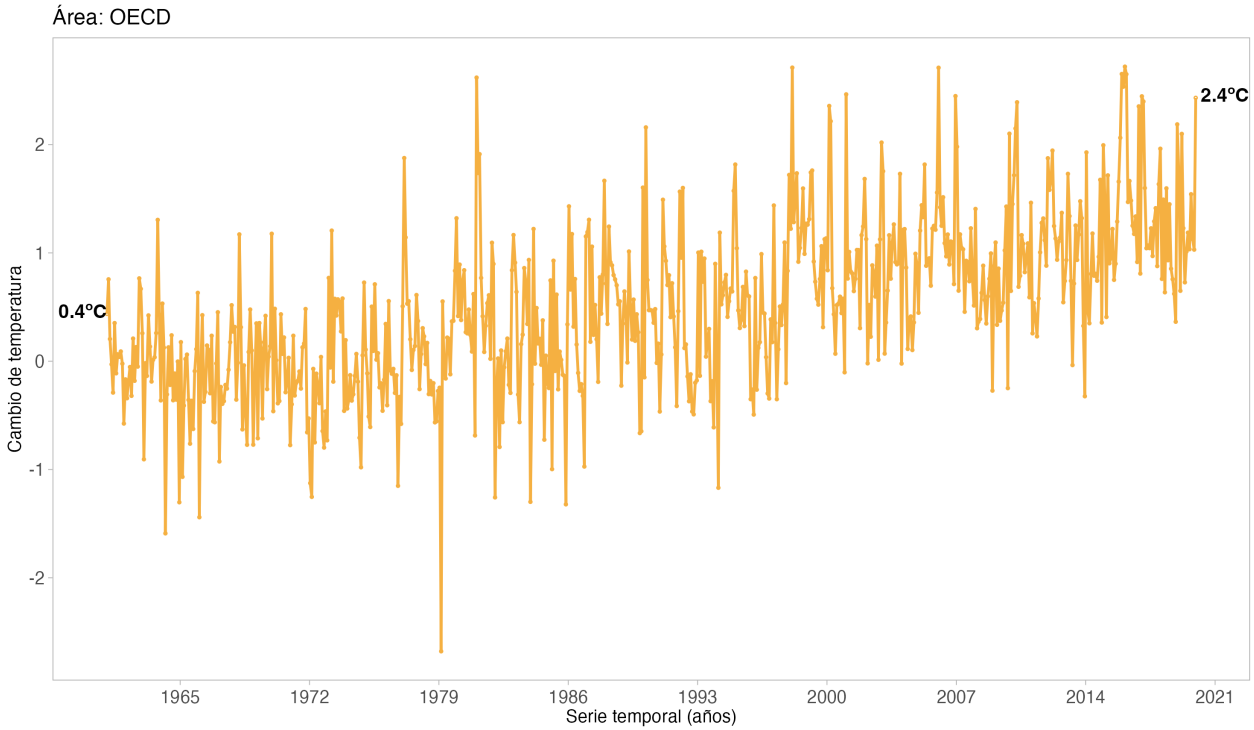
```
fun(areas="World", desde=1961, hasta=2019)
```

```
   area  months year temperature_change       date
1 World January 1961              0.399 1961-01-15
2 World January 1962              0.104 1962-01-15
3 World January 1963             -0.171 1963-01-15
4 World January 1964             -0.311 1964-01-15
5 World January 1965              0.206 1965-01-15
6 World January 1966              0.458 1966-01-15
```



```
fun(areas="OECD", desde=1961, hasta=2019)
```

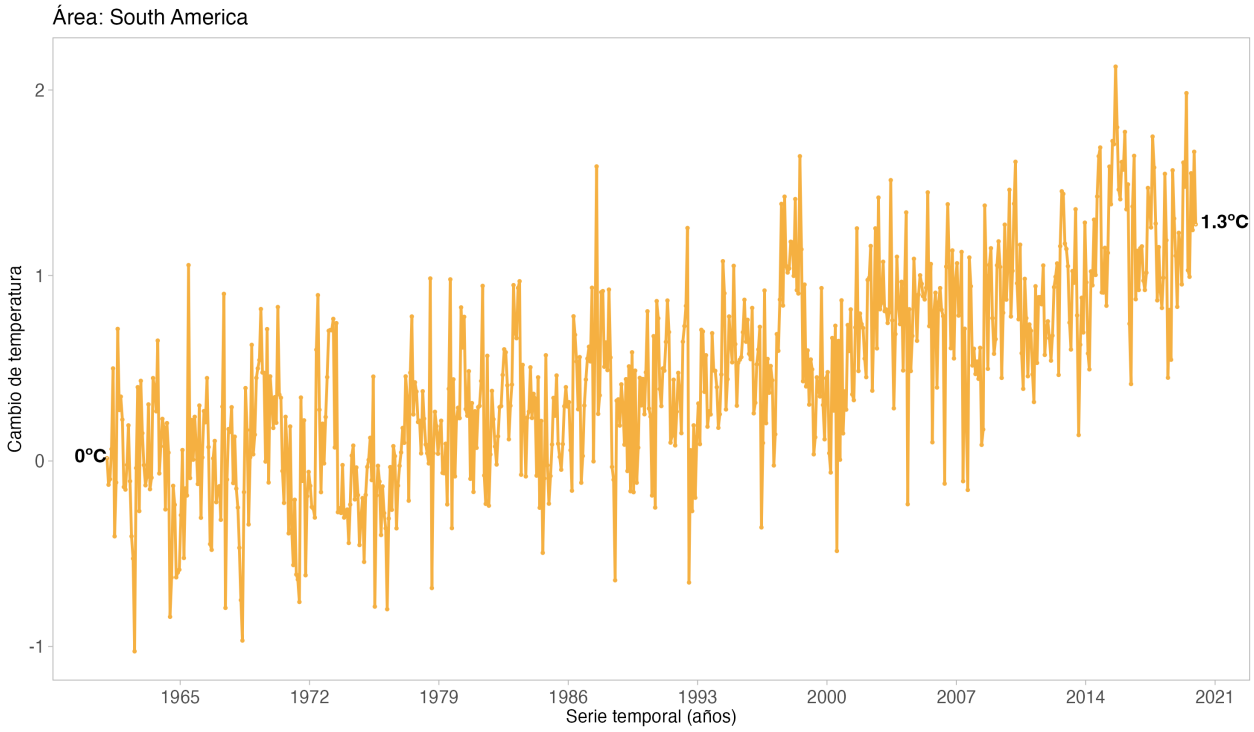
```
  area  months year temperature_change       date
1 OECD January 1961              0.437 1961-01-15
2 OECD January 1962             -0.168 1962-01-15
3 OECD January 1963             -0.906 1963-01-15
4 OECD January 1964              0.532 1964-01-15
5 OECD January 1965              0.176 1965-01-15
6 OECD January 1966             -1.441 1966-01-15
```



Elaboración propia a partir de los datos de cambio de temperatura entre 1961 - 2019

```
fun(areas="South America", desde=1961, hasta=2019)
```

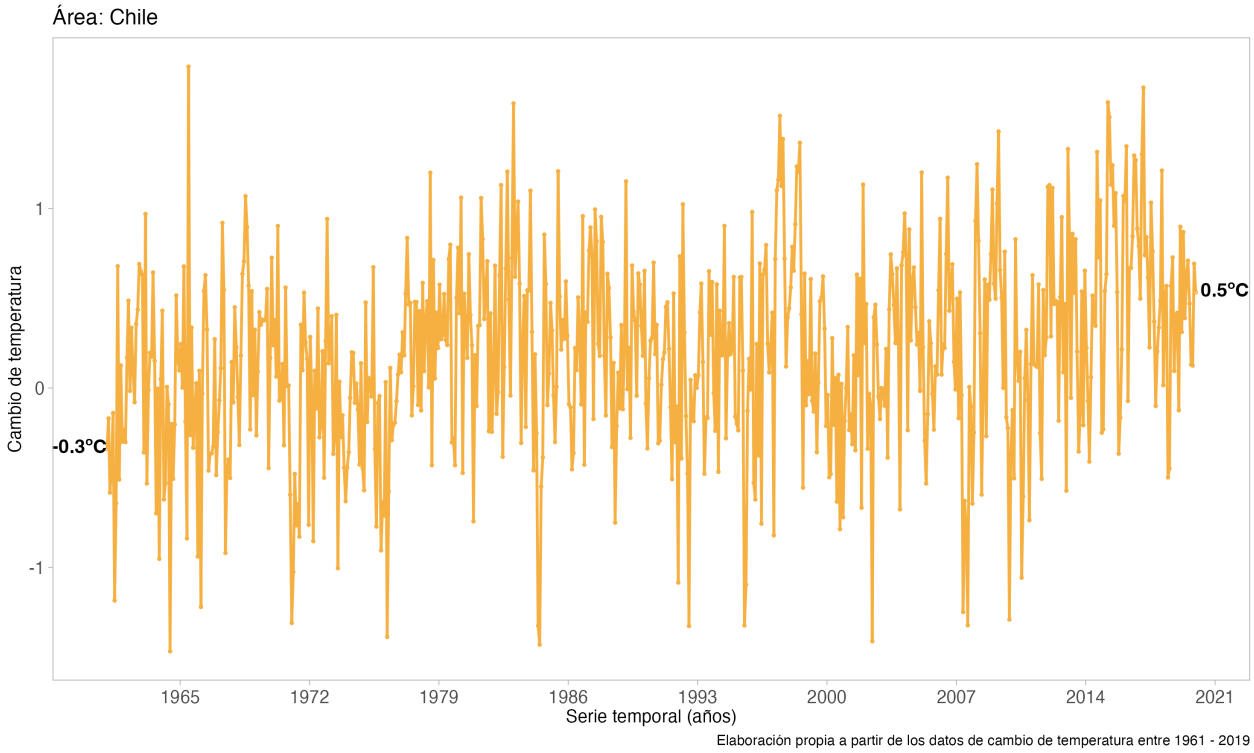
```
           area  months year temperature_change       date
1 South America January 1961              0.015 1961-01-15
2 South America January 1962             -0.154 1962-01-15
3 South America January 1963             -0.022 1963-01-15
4 South America January 1964              0.227 1964-01-15
5 South America January 1965             -0.292 1965-01-15
6 South America January 1966              0.299 1966-01-15
```



Elaboración propia a partir de los datos de cambio de temperatura entre 1961 - 2019

```
fun(areas="Chile", desde=1961, hasta=2019)
```

```
   area  months year temperature_change       date
1 Chile January 1961              -0.340 1961-01-15
2 Chile January 1962              -0.301 1962-01-15
3 Chile January 1963              -0.359 1963-01-15
4 Chile January 1964               0.431 1964-01-15
5 Chile January 1965               0.245 1965-01-15
6 Chile January 1966               0.096 1966-01-15
```



Las interpretaciones pueden considerar aspectos como la **(1) tendencia general en el cambio de temperatura** y **(2) patrones específicos en el tiempo**.