

Desafío R

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-área 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:
- Limpiar los nombres de las columnas para que queden en minúsculas. *Hint:* Revise las funciones `tolower()` o `str_to_lower()` (2 puntos).
 - 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).
 - 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 funciones `mutate()`, `str_to_replace()` (5 puntos).
 - 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 unirlos 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(area="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):

