Taller 2 y 3: Manejo y Visualización de Datos

Programación en R

Profesor Santiago Tellez Cañas

Fecha de entrega: Domingo, Mayo 2 de 2021

- Este taller lo deben resolver individualmente y lo deben entregar a más tardar el domingo 2 de mayo a las 11:59 pm.
- **Todas las gráficas incluidas en este taller deben tener un título. Las etiquetas de los ejes x y y deben cambiarse para incluir etiquetes que den cuenta de la información contenida en cada eje. Si hay leyendas, también se debe cambiar el título de la leyenda. Además, deben usar una plantilla diferente a la que por defecto usa ggplot. Igualmente, los colores de puntos, líneas, etc, deben ser distintos a los que por defecto usa ggplot. Las escalas de los ejes deben ser consistentes con la naturaleza de la información que se está mostrando y deben poder verse adecuadamente.**
- **Las preguntas 3 a 15 valen 2,6. Las preguntas 16 a 17 valen 1,2. La pregunta 18 vale 1,2.**
- Estaré atento a cualquier pregunta que tengan por correo electrónico y en el horario de atención. El miércoles 21 de abril no tendré atención de 4 a 5. Entonces estaré el viernes 23 de abril de 10 a 12, el miércoles 28 de abril de 4 a 5 y el viernes 30 de abril de 10 a 12.

1. Cree un repositorio de Github que se llame Taller2_Apellido. Use su primer apellido en reemplazo de la palabra Apellido en el nombre. Este repositorio debe ser privado. Después de crearlo, cree un proyecto en RStudio y vincúlelo al repositorio que creó.
2. Dentro de este proyecto, cree un notebook en el que va a desarrollar el taller.
3. Use la siguiente función para descargar de la API del Banco Mundial la información de los indicadores NY.ADJ.NNTY.PC.KD, que contiene el ingreso nacional percapita neto ajustado, y SP.DYN.LE00.IN que contiene la expectativa de vida al nacer para ambos sexos.

```
datos_bm <- wb_data(indicator = c("NY.ADJ.NNTY.PC.KD", "SP.DYN.LE00.IN"),
                    start_date = 2000, end_date = 2020,
                    return_wide = FALSE)
```

4. Use la siguiente función para descargar la tabla de información disponible para los países:

```
paises_bm <- wb_countries()
```

5. En la base países_bm seleccione las siguientes variables:iso3c, region, income_level.

6. Una las bases de datos datos_bm y países_bm.

7. En la variable indicator_id cambie los valores NY.ADJ.NNTY.PC.KD y SP.DYN.LE00.IN por ing_nac_ajustado y expectativa_vida, respectivamente.

8. Use la función `pivot_wider` para transformar la base de datos de manera que se consideren datos limpios (i.e. *tidy data*). En la base resultante deben haber 8 variables: indicator, iso2c, iso3c, country, date, region, income_level, ing_nac_ajustado y expectativa_vida. Antes de correr la función revise para qué sirve y cómo se usa el argumento id_cols dentro de esta función.

9. Reorganice las variables para que aparezcan las siguientes al comienzo de la base de datos: region, income_level, country, date, ing_nac_ajustado y expectativa_vida.

10. Usando la base anterior, cree una base de datos que solamente contenga las variables country, date, region, income_level, ing_nac_ajustado y expectativa_vida. Esta base debe llamarse bm_principales. Use alguna función del paquete naniar para explorar los valores faltantes en esta base. ¿Cuál variable parece tener mayor número de valores faltantes?

11. Use la función `gg_miss_fct` para mostrar el comportamiento de los valores faltantes en esta base de datos, de acuerdo con las categorías de la variable region. ¿En cuál región parecen haber más valores faltantes para las variables expectativa_vida e ing_nac_ajustado? Repita el ejercicio pero ahora usando la variable income_level. ¿En cuál nivel de ingreso parecen haber más valores faltantes para las variables mencionadas?

12. Filtre el año 2015 y realice un diagrama de dispersión de las variables expectativa_vida (eje y) e ing_nac_ajustado (eje x) para ese año. El color de los puntos debe depender de la variable region. ¿Qué muestra la gráfica sobre la relación entre las variables para el 2015?

13. Filtre la información de Colombia y realice un diagrama de lineas que muestre la evolución de expectativa_vida en el periodo de análisis. ¿Qué muestra la gráfica sobre la evolución de la expectativa de vida desde el 2000?

14. Filtre los años 2000 y 2015 y realice una densidad de la variable expectativa_vida en la que se muestre de un color distinto la distribución para cada uno de esos dos años. ¿Qué diferencias se puede ver en la distribución de expectativa_vida para estos dos años?

15. Filtre la información para el año 2015 y realice una gráfica que muestre la densidad de la variable expectativa_vida de manera separada para cada región. Use las funciones facet_grid o facet_wrap. ¿Qué diferencias importantes se encuentran en la distribución de expectativa_vida entre las regiones?

16. Descargue los microdatos de la Gran Encuesta Integrada de Hogares para el mes de febrero de 2021. Use un loop for para importar los datos que corresponden a los módulos cabecera y resto. Para los módulos de cabecera, una el módulo de características generales (personas) y el de vivienda y hogares. Después, use un loop for para unir los módulos restantes. Repita el mismo ejercicio para los módulos de resto. Cuando haya terminado, una cabecera y resto usando la función `bind_rows`. Al final, tendrá la información nacional de la GEIH para el mes de febrero de 2021.

17. Usando la base resultante en el punto anterior, emplee la función `mutate`, en conjunto con la función `across` para convertir en factor las variables P6250, P6020, P6440, P7310, y P7430 (Revisen el diccionario de datos para ver el contenido de cada una de estas variables). En cada una de estas variables la categoría 1 debe tener la etiqueta Sí, y la categoría 2 debe tener la etiqueta No.

18. 1. Importen la base de datos de contagios de la Covid 19 para Colombia. Pueden usar la siguiente función:

```
covid_19 <- read_csv("https://www.datos.gov.co/api/views/gt2j-8ykr/rows.csv?accessT
```

    2. Cambien el nombre a la variable fecha reporte web a fecha_reporte.
    3. Asegúrese de que la variable fecha_reporte es identificada por R como una fecha.
    4. Use la función `summarize` o la función `count` en conjunto con la función `group_by` para contar el número de casos que se ha reportado en cada fecha de reporte.
    5. Realice un diagrama de líneas para mostrar la evolución del número de casos diarios en Colombia durante la pandemia.
    6. [Bono por 0,5] Cree una variable que contenga para cada día, el promedio móvil de casos de los 7 días anteriores. Realice el mismo diagrama de líneas pero usando la variable que acaba de crear.

19. Agrégueme como colaborador en el repositorio de Github y publique el link del repositorio, junto con la versión html del notebook en el Aula Virtual, en el espacio que crearé para ello.