Taller de R: Estadística y Programación

Taller para evaluar conocimientos módulos 1 y 2

15/09/2021

En este taller se evalúan los módulos 1 y 2 del curso. Se presentan 2 tipos de taller (A y B), pero usted solo debe desarrollar 1 de ellos. Sea creativo en su código (no hay una respuesta única, todos los métodos que permitan obtener la misma respuesta son validos). Cuando encuentre una ayuda en línea que le permita solucionar algún problema, no olvide citar la fuente. Por último, lea atentamente las instrucciones del taller.

Instrucciones

- No seguir las instrucciones tiene una penalización del **10%** de la nota total.
- Debe crear un repositorio que contenga las carpetas que están en este repositorio clik aquí y debe invitar como colaborador del repositorio al usuario **eduard-martinez** de github.
- Debe poner en el archivo de excel del equipo de Tams (Documents/General/task/task_1/task_1.xls) su código estudiantil y el link al repositorio con la respuesta de su taller.
- El taller debe ser terminado antes de las 23:59 horas del 01 de octubre de 2021.
- Por favor sea lo más organizado posible y comente paso a paso cada línea de código, pero recuerden **NO** usar ningún acento o carácter especial dentro del código para evitar problemas al abrir los script en los diferentes sistemas operativos.
- En las primeras líneas del script debe escribir su nombre y la versión de R sobre la que está trabajando.
- Recuerde que debe elegir y desarrollar solo 1 taller (A o B).

Sobre la GEIH

La recolección de la Gran Encuesta Integrada de Hogares -GEIH- empezó el 7 de agosto de 2006 en su módulo central de mercado laboral e ingresos y, a partir del 11 de septiembre, con su módulo de gastos de los hogares. A partir del 10 de julio de 2006 se amplió la cobertura de la ECH a once ciudades adicionales, a las trece principales ciudades y áreas metropolitanas, al resto de cabeceras y al resto rural; cobertura que en la actualidad mantiene la GEIH. Actualmente la encuesta se ha especializado en la medición de la estructura del mercado laboral y los ingresos de los hogares. Esta tiene una muestra total anual de 240.000 hogares aproximadamente, lo que hace que sea la de mayor cobertura a nivel nacional.

La GEIH recoge información a tres niveles geográficos: Áreas metropolitanas, Cabeceras y Restos. En `Areas` se recoge información para las 13 principales áreas metropolitanas del país. Por su parte, `cabecera` lo hace para todas las cabeceras municipales (o zonas urbanas del país, inclusive las áreas metropolitanas). Finalmente, `resto` recoge información para las zonas rurales del país. Para cada nivel geográfico se puede acceder de manera libre a los siguientes módulos:

- 1. Características generales personas: se recoge información de algunas características observables de las personas como la edad, sexo, ...
- 2. Desocupados: información de las personas que reportaron estar desocupadas pero que se encontraban buscando empleo.

- 3. Fuerza de trabajo: información de las personas que pertenecen a la fuerza de trabajo.
- 4. Inactivos: información de las personas que reportaron no estar trabajando pero que tampoco están buscando empleo.
- 5. Ocupados: información de las personas que reportaron estar ocupadas al momento de la encuesta.
- 6. Otras actividades y ayudas en la semana: información sobre ingresos.
- 7. Otros ingresos: información sobre ingresos.
- 8. Vivienda y hogares: características de la vivienda y el hogar de la persona encuestada.

Todos los módulos poseen estas 3 variables (`secuencia_p` , `orden` y `directorio`) que permiten cruzar información de todos los módulos para un mismo individuo. Puede obtener una descripción detallada de todas las variables aquí. Sin embargo, para los propósitos de este taller puede ser suficiente con la información que le será suministrada en los siguientes incisos.

Taller A

1. Vectores

Cree un vector que contenga los números del 1 al 100, posteriormente cree otro vector que contenga los números impares de 1 a 99. Use el vector de números impares para crear un vector con los números pares de del primer vector.

2. Limpiar una base de datos

Importe la base de datos **cultivos** que se encuentra en la carpeta *data/input*, limpie la base de datos eliminando las observaciones que no tienen información relevante. Luego pivotee la base de datos para que quede en formato long.

3. GEIH

3.1. Importar

Importe las bases de datos **Cabecera - Caracteristicas generales (Personas).rds**, **Cabecera - Ocupados.rds**, **Cabecera - Desocupados.rds**, **Cabecera - Inactivos.rds** y **Cabecera - Fuerza de trabajo.rds** de la carpeta **data/input/2019**. Luego use las variables `secuencia_p`, `orden` y `directorio` para unir en una única base de datos. Mantenga en la base de datos únicamente las variables `secuencia_p`, `orden`, `directorio`, `P6020, P6040, P6030S1, P6440, P6450, P6920, INGLABO, DPTO, fex_c_2011, ESC, MES, P6050` y las variables que usted va a generar para saber si una persona es ocupada, desocupada, inactiva y/o fuerza de trabajo.

3.2 Decsriptivas

Use las funciones `ggplot()`, `group_by()` y `summarize()` entre otras, para generar algunas estadísticas descriptivas (gráficos y tablas) numero de ocupados, desocupados y los ingresos laborales promedio. Tenga en cuenta algunas dimensiones como departamento, sexo y edad. Las tablas las puede plotear sobre la consola, pero los gráficos los debe exportar en formato `.jpeg` a la carpeta *views*. Debe generar almenos 5 gráficos y 5 tablas.

Hint: Puede validar algunos de sus resultados en el visor GEIH del DANE.

Taller B

1. Organizar GEIH

1.1. Importar bases de datos

Importe a R los archivos contenidos en las carpetas *data/input/2019* y *data/input/2020*. Para los módulos de *ocupados, inactivos, desocupados* y *fuerza de trabajo* asegúrese de crear una variable categórica que le permita identificar si las personas entrevistas están en una de las categorías mencionadas anteriormente.

Hint: para evitar duplicados de algunas variables, de cada modulo deje únicamente las variables `secuencia_p` , `orden` y `directorio`. De los demás módulos deje únicamente las variables `P6020, P6040, P6030S1, P6440, P6450, P6920, INGLABO, DPTO, fex_c_2011, ESC, MES` y `P6050`. Además, para no generar 20 objetos puede intentar guardarlos en 2 listas una por cada año, asegurándose no perder el orden de los archivos.

1.2. Unir datos

Use algunas funciones del paquete tidyverse que le permitan unir (por filas y columnas) en dos bases de datos que contenga todos los módulos de `cabecera` y `resto` respectivamente. Limpie la consola y deje sobre el entorno de R únicamente estos dos objetos.

Hint: Asegúrese de crear una variable que le permita identificar las observaciones de cada año.

1.3 Una base nacional

Cree un objeto llamado `nacional` que contenga las bases de datos de cabecera y los datos de resto.

Hint: Asegúrese de crear una variable que le permita identificar las observaciones urbanas (`cabecera`) y las rurales (`resto`). **Hint:** Use la función `skim` para describir la base de datos.

1.4 Descriptivas

Use las funciones `ggplot()`, `group_by()` y `summarize()` entre otras, para generar algunas estadísticas descriptivas (gráficos y tablas) numero de ocupados, desocupados y los ingresos laborales promedio. Tenga en cuenta algunas dimensiones como año, departamento, sexo, urbano/rural y edad. Las tablas las puede plotear sobre la consola, pero los gráficos los debe exportar en formato `.jpeg` a la carpeta *views*.

Hint: Puede validar algunos de sus resultados en el visor GEIH del DANE.