Taller de R: Estadística y Programación

Taller 1

15/09/2020

En este taller se evaluan los módulos 1 y 2 del curso. Se presentan 2 tipos de taller (A y B), pero usted solo debe desarrollar 1 de ellos. Para realizar este trabajo, podrán trabajar de manera individual o hacer grupos de hasta dos personas. Sea creativo en su código (no hay una respuesta unica, todos los métodos que permitan obtener la misma respuesta, son validos). Cuando encuentre una ayuda en linea que le permita solucionar algun problema, no olvide dar los creditos. Por ultimo lea atentamente las instrucciones del taller.

Instrucciones

- No seguir las instrucciones tiene una penalización del **10%** de la nota total.
- Este taller representa el 30% de la nota del curso.
- El taller estará disponible en SICUA desde las 23:00 horas del martes 15 de septiembre de 2020.
- Debe elegir y desarrollar solo 1 taller (A o B).
- Por favor sea lo más organizado posible y comente paso a paso cada linea de código, pero recuerden **NO** usar ningún carácter especial dentro del código para evitar problemas al abrir los script en los diferentes sistemas operativos.
- El taller debe ser entregado antes de las 23:59 horas del domingo 4 de octubre de 2020.
- Si usted decide hacer el taller A, debe enviar el script con la solución al correo ef.martinezg@uniandes.edu.co con el siguiente asunto: *Taller 1 de R, 2020-2*. Al comienzo del script debe estar el nombre(s) y el código de la(s) persona(s) que trabajaron en el código. Además el script debe ser guardado con el primer apellido de cada uno de los integrantes más el código estudiantil separado por _ , tal como se muestra a continuación: *martinez_201725842_arevalo_201522054.R*.
- Si usted decide hacer el taller B, debe invitar al usuario eduard-martinez como colaborador de su repositorio. Este vídeo le puede ayudar a entender mejor el funcionamiento de los repositorios de GitHub.

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 propositos de este taller puede ser suficiente con la información que le será suministrada en los siguientes incisos.

Taller A

1 Organizar la GEIH

1.1. Importar bases de datos

Importe a R los archivos contenidos en las carpetas *data/original/junio 2019* y *data/original/junio 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`.

Bonus: 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 las funciones `merge` y `rbind.fill` para crear dos bases de datos que contenga todos los modulos 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.

Bonus: Intente hacerlo combinando las funciones `ls()`, `grepl()` y `rm()`.

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`).

1.4 Descriptivas

Use las funciones `ggplot()`, `group_by()` y `summarize()` entre otras, para generar algunas estadísticas descriptivas (gráficos y tablas) de la tasa de ocupación, la tasa de desempleo y los ingresos laborales. Tenga en cuenta algunas dimensiones como año, sexo, urbano/rural, tipo de contrato (P6440) y la edad. Las tablas las puede plotear sobre la consola, pero los graficos los debe exportar en formato `.jpeg` a la carpeta *results*.

Hint: En la carpeta *help* usted encuentra un pdf del DANE que le ayudara a calcular las tasas de desempleo y ocupación. De igual forma, puede validar algunos de sus resultados en el visor GEIH del DANE.

2.Reshape

De la carpeta *data/original* importe la base de datos llamada \textit{tasa_deso_sexo.xlsx}. Use las funciones `fill`, las funciones del paquete `reshape2` y las demás funciones que usted considere necesarias para convertir esa base de datos en la base de datos *data/procesada/tasa ocupados por sexo.rds*

Taller B

En este taller usted debe usar los datos de la GEIH para los años 2019 y 2020 que se encuentran disponibles en la carpeta *data/original* para responder las siguientes preguntas:

• Hay una brecha de genero en la tasa de ocupación y la tasa de desempleo en Colombia? En términos de ingresos laborales? • Quienes han perdido el mayor numero de empleos de junio de 2019 a junio de 2020? Ha sido igual para diferentes niveles de educación, tipo de contrato o ubicación geográfica?

Use gráficos y tablas que le permitan sustentar sus respuestas. Cree un repositorio (privado) en GitHub para compartir sus resultados. El repositorio debe contener las siguientes carpetas: codes, data y results. Debe incluir sus respuestas en en el archivo README.md del respositorio. Este vídeo le puede ayudar.