

Universidad Nacional de Santiago del Estero

Programa de posgrado “Métodos y técnicas para la investigación aplicada a la realidad socioeconómica regional”

Curso Introducción a la Ciencia de Datos para Científicos Sociales

Trabajo final

Consideraciones preliminares

El trabajo deberá ser realizado en **grupos de hasta 4 personas**. Todas las consignas deben ser resueltas usando los verbos del paquete dplyr (select(), filter(), mutate(), arrange(), summarize(), etc.) y otros paquetes del ecosistema tidyverse (ggplot2, lubridate, etc.).

Si tiene que generar objetos intermedios, por ejemplo, tablas de resultados intermedios, hacerlo con nombres que sean lo más descriptivos posible de su contenido. Es posible que para resolver alguna de las consignas sea necesario consultar la documentación de algunas de las funciones. Recuerde que puede hacerlo con el comando help([nombre_funcion])

Algunos sitios útiles

- <https://ggplot2.tidyverse.org/>
- <https://dplyr.tidyverse.org/>
- <https://lubridate.tidyverse.org/>
- <https://forcats.tidyverse.org/> (paquete sumamente útil para trabajar con variables tipo factor())
- <https://bitsandbricks.github.io/ciencia-de-datos-gente-sociable/>
- <http://www.indec.gov.ar>
- <https://r4ds.had.co.nz/> R For Data Science (manual –versión inglés-)
- <https://es.r4ds.hadley.nz/> R For Data Science (manual –versión castellano-)

Fecha de pre-entrega: 15/11/2019

Fecha de entrega: 09/12/2019

Materiales a entregar

- Un archivo Rscript en el que se dispone el código que resuelve las consignas. El mismo deberá contener el código completo desde la importación de los archivos hasta el resultado final, pasando por todos los objetos intermedios y procesamientos que se generen. Recordar que para comentar porciones del código (por ejemplo, para
- Un archivo .docx que contiene las salidas respuestas de cada consigna y, en caso de ser necesario, la interpretación de las mismas.

Ambos archivos deberán estar nombrados de la siguiente forma:

[apellido]_tpfinal.r

[apellido]_tpfinal.docx

Dudas, consultas, etc.: german.rosati@gmail.com

Consignas

Se provee el archivo adjunto (Individual_t414.zip) que contiene comprimido un archivo .csv con las respuestas individuales de la Encuesta Permanente de Hogares del 4to. trimestre del año 2014. A partir del mismo resolver las siguientes consignas:

1. Calcular las tasas de actividad, empleo y desempleo según sexo, para jóvenes entre 18 y 35 años. El resultado final debe ser una tabla o dataframe que contenga los tres indicadores. ¿Qué puede observar en las distribuciones obtenidas?

Nota: recordar las siguientes definiciones

- Tasa de actividad = $(\text{Población ocupada} + \text{Población desocupada}) / \text{Población total} * 100$
- Tasa de empleo = $\text{Población ocupada} / \text{Población total} * 100$
- Tasa de desempleo = $\text{Población desocupada} / \text{Población activa}$

2. Calcular el salario promedio por sexo, para dos grupos de edad: 18 a 35 años y 36 a 70 años. ¿Qué conclusión puede extraer de los datos obtenidos?

Nota: La base debe filtrarse para contener únicamente OCUPADOS ASALARIADOS

3. Graficar la distribución del ingreso por ocupación principal según categoría ocupacional. Generar un histograma y un boxplot. Haga una breve interpretación de los resultados.

4. ¿Existe alguna relación entre el ingreso (p21) y la edad de las personas? ¿Y entre el ingreso y el nivel educativo? Generar los gráficos que le permitan responder a ambas preguntas e interprete los resultados.