INSTITUTO DE SOCIOLOGÍA
FACULTAD DE CIENCIAS SOCIALES

SOL4001
Procesamiento Avanzado de Bases de Datos en R

Mauricio Bucca (mebucca@uc.cl)

CURSO	: **Procesamiento Avanzado de Bases de Datos en R**
NOMBRE INGLÉS	: Advanced Database Processing using R
SIGLA	: SOL-4001
CRÉDITOS	: 5 UC / 2 SCT
PROFESOR	: **Mauricio Bucca**, Sociólogo de la Universidad Católica de Chile. Magíster y Doctor en Sociología, Cornell University, (Estados Unidos).
AYUDANTES.	: Martín Aranzaes (maaranzaes@uc.cl), Estudiante de Magíster en Sociología, Universidad Católica de Chile. Matías Deneken (m.deneken@uc.cl), Estudiante de Magíster en Sociología, Universidad Católica de Chile. Sebastián Urbina (saurbina@uc.cl), Magíster en Sociología, Universidad Católica de Chile.
FECHAS	: Desde el 30 de agosto al 17 de noviembre del 2022
HORARIO	: Martes y Jueves de 18:00 a 20:00 hrs.
LUGAR	: Online, plataforma Zoom.

I. DESCRIPCIÓN

Este curso aborda aspectos avanzados en el procesamiento de bases de datos, tales como manejo de variables, consolidación de bases de datos, buenas prácticas de programación y producción de reportes automatizados y replicables. Al final del curso se espera que los alumnos puedan analizar bases de datos de mediana a avanzada complejidad. El desarrollo de los contenidos será en el programa R, un software estadístico gratuito y de código abierto que se encuentra entre los más utilizados en ámbitos académicos e investigación aplicada.

II. OBJETIVOS

- Desarrollar destrezas en la creación, importación, exportación, fusión y modificación de bases de datos.
- Entrenarse en el cálculo de indicadores para diferentes unidades de análisis en una misma base de datos.
- Adquirir habilidades para llevar a cabo análisis de datos de modo efectivo, eficiente y reproducible.

III. CONTENIDOS

1. Principios de programación en el lenguaje R
2. Limpieza, manipulación y validación de datos
3. Automatización de análisis, resultados y reportes
4. Organización y documentación de análisis de datos

IV. METODOLOGÍA

Este curso se desarrollará en modalidad online y utilizará las siguientes herramientas pedagógicas:

- Clases sincrónicas: Clases expositivas en directo online, vía *streaming*, una vez por semana; Discusión de textos y aprendizaje basado en problemas; Trabajos aplicados y breves presentaciones en clase de los estudiantes. Las clases sincrónicas serán de dos tipos
    - Clases teóricas dictadas por el profesor a la totalidad de la clase
    - Clases prácticas dictadas por los ayudantes. Cada ayudante trabajará con subgrupo fije de estudiantes (1/3 de la clase).
- Clases a-sincrónicas: 8 horas de clases expositivas y/o tutoriales disponibles a través de videos pre-grabados.

Todas las clases en vivo serán grabadas y estarán disponibles por 7 días corridos, a través de un enlace en la plataforma web Classroom.

Las cápsulas formativas se subirán a la plataforma Classroom a medida que avancen los contenidos del curso y permanecerán disponibles hasta el término del curso.

La primera cápsula disponible explicará cómo descargar e instalar R y RStudio. Esta primera cápsula estará disponible antes del inicio del curso de tal manera que los estudiantes puedan comenzar la primera clase con R y RStudio ya instalados en sus computadores personales.

V. EVALUACIÓN

La nota final del curso se calcula a partir de dos componentes de evaluación: Tareas (60%) y Trabajo final (40%).

Tareas (60%)

El componente Tareas consiste en el desarrollo de 5 reportes que ponderan en total **60% de la nota final del curso**. Esta actividad se desarrolla en forma individual. Las fechas de publicación y entrega de cada una de las tareas están especificadas en el programa detallado del curso.

- Tarea #1: Se asignará el día 01 de septiembre.
- Tarea #2: Se asignará el día 22 de septiembre.
- Tarea #3: Se asignará el día 13 de octubre.
- Tarea #4: Se asignará el día 20 de octubre.
- Tarea #5: Se asignará el día 11 de noviembre.

Trabajo final (40%)

El trabajo final pondera un **40% de la nota final del curso**, y se desarrolla en forma individual. Se asignará el día 15 de noviembre.

VI. INTEGRIDAD ACADÉMICA

Se espera que los alumnos mantengan altos estándares de integridad académica. Los casos de plagio o copia durante la aplicación de alguna evaluación o trabajo serán sancionados con un 1.0 y serán informadas obligatoriamente a la subdirección de educación continua. Otras posibles infracciones a la honestidad académica también serán derivadas a la subdirección donde se evaluarán posibles sanciones (ver Reglamento del Alumnos de Educación Continua).

Las peticiones de recorrección deberán hacerse por escrito al profesor en un plazo de máximo 5 días hábiles desde la entrega de las evaluaciones. La solicitud de recorrección deberá estar debidamente fundamentada.

VII.BIBLIOGRAFÍA

El curso es auto-contenido y **no completa lecturas obligatorias**. No obstante, en la presentación de cada clase se sugerirán lecturas para reforzar y complementar lo aprendido.

Análisis de datos y programación en R:

Básicos (disponibles en la UC):

- Hadley Wickham (2009), ggplot2 Elegant Graphics for Data Analysis. Springer
- Bradley C. Boehmke (2016), Data Wrangling with R. Springer
- Robert Kabacoff (2015), R in Action Data Analysis and Graphics with R. Manning Publications
- Keon-Woong Moon (2016), Learn ggplot2 Using Shiny App. Springer
- Matt Wiley, Joshua F. Wiley (2016), Advanced R. Data Programming and the Cloud. Apress.

Otros:

- Hadley Wickham (2015) Advanced R, CRC Press, Taylor & Francis Group, Boca Raton, FL.
- Hadley Wickham and Garrett Grolemund (2017). R for Data Science. Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, Inc.,
- Garrett Grolemund (2014). Hands-On Programming with R. O'Reilly Media, Inc.,
- Chris Beeley (2013). Web Application Development with R Using Shiny. Packt Publishing.
- Winston Chang (2013). R Graphics Cookbook. O'Reilly Media, Inc.,
- Yihui Xie (2013). Dynamic Documents with R and knitr. O'Reilly Media, Inc.,

VIII. PROGRAMA DETALLADO DE CLASES

Clase 01 y 02: Introducción a R y Rstudio. Sintaxis y operaciones básicas.

30 de agosto

1. Interfaz R y RStudio
2. Manejo de archivos
3. Operaciones matemáticas
4. Operaciones lógicas
5. Operaciones de vectores
6. Otras funciones útiles
7. Manejo de librerías

Clase 03 y 04: Introducción a bases de datos

01 y 06 de septiembre

1. Creación y manipulación de bases de datos
2. Variables e individuos
3. Extracción, modificación de variables, creación de nuevas variables
4. Importación y exportación de bases de datos
5. Operaciones básicas

Clase 05 y 06: Workflow

08 y 13 de septiembre

1. Uso de Scripts
2. Buenas prácticas de programación
3. Construcción de un "workflow" efectivo y ordenado
4. Exportación de resultados
5. Replicabilidad

Clase 07: tidyverse, pipes y funciones básicas dplyr

15 y 20 de septiembre

1. Librería dplyr
2. Ordenamiento de bases de datos
3. Filtro de bases de datos
4. Selección de variables

Clase 08 y 09: Creación y transformación de variables con dplyr

22 y 27 de septiembre

1. Creación de nuevas variables (indicadores y variables de estratificación)
2. Recodificación de variables
3. Cálculo de variables en diferentes unidades de medición

Clase 10 y 11: Resumen de datos agrupados y combinación de bases de datos con dplyr

29 de septiembre y 04 de octubre

1. Estadísticas básicas
2. Resumen de las variables en una base de datos
3. Resumen de variables por grupos
4. Juntar bases de datos con una llave común

Clase 12 y 13: Transformación de datos anchos y largos con tidyr

06 y 11 de octubre

1. Concepto de bases de datos ordenadas (tidy).
2. Transformación de datos "anchos" y "largos"

Clase 14 y 15: Tratamiento de datos faltantes con tidyr

13 y 18 de octubre

1. Manipulación de datos faltantes con funciones bases de R
2. Herramientas de tidyverse para manipulación de datos faltantes

Clase 19 y 20: Iteración y automatización con purrr

20 y 25 de octubre

1. Funciones personalizadas
2. Herramientas de iteración
3. "Functional programming"

Clase 16 y 17: Visualización de datos con ggplot2

27 de octubre

1. La "gramática" de ggplot
2. Gráficos para una sola variable
3. Gráficos para relaciones entre variables

Clase 18: Visualización de datos con ggplot2

03 y 08 de noviembre

1. Personalización de gráficos
2. Gráficos por grupo
3. Ejemplos de gráficos avanzados
4. Exportación de figuras

Clase 21 y 22: Reportes automatizados

10 y 15 de noviembre

1. Rmarkdown y librería knirt
2. Escritura de reportes automatizados y replicables
3. Un primer acercamiento a presentaciones automatizadas en Xaringan

Clase 23: Recapitulación

17 de noviembre

1. Uso avanzado de Scripts

2. Construcción de un “workflow” efectivo, ordenado y automático
3. Exportación avanzada de resultados
4. Replicabilidad
5. Panorámica de otras herramientas útiles
6. Editores de texto

Cápsulas Formativas

El curso contempla la entrega de material formativo complementario a través de 8 horas de cápsulas formativas disponibles en videos pre-grabados que pueden ser revisados por los alumnos en cualquier momento durante el desarrollo del curso.