



SOL4001

Procesamiento Avanzado de Bases de Datos en R

Mauricio Bucca (mebucca@uc.cl)

27 de julio de 2020

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).
AYUDANTE.	: Martín Aranzaes, Estudiante de Magister, Universidad Católica de Chile.
FECHAS	: Desde el 26 de septiembre al 12 de diciembre del 2020.
HORARIO	: Jueves de 18:00 a 20:50 hrs. y Sábados de 9:00 a 12:00 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.
- Clases a-sincrónicas: clases expositivas y/o tutoriales disponibles a través de videos pre-grabados.

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 o en grupos de máximo dos personas. 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 Jueves 01 de octubre de 2020
- Tarea #2: Se asignará el día Jueves 08 de octubre de 2020
- Tarea #3: Se asignará el día Jueves 22 de octubre de 2020
- Tarea #4: Se asignará el día Jueves 07 de noviembre de 2020
- Tarea #5: Se asignará el día Jueves 19 de noviembre de 2020

Trabajo final (40%)

El trabajo final pondera un **40% de la nota final del curso**, y se desarrolla en forma individual. Se asignará a más tardar el día Jueves 26 de noviembre de 2020.

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 corrección deberán hacerse por escrito al profesor en un plazo de máximo 5 días hábiles desde la entrega de las evaluaciones. La solicitud de corrección 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 Grolmund (2017). R for Data Science. Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, Inc.,
- Garrett Grolmund (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: Introducción a R y Rstudio. Sintaxis y operaciones básicas.

Sábado 26 de septiembre, 9:00 a 12:00 hrs.

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

Instructor: Mauricio Bucca

Clase 02: Introducción a bases de datos

Jueves 01 de octubre, 18:00 a 20:00 hrs.

1. Creación 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

Instructor: Mauricio Bucca

Clase 03: Taller #1, Primer acercamiento a bases de datos en R (CAsen 2017)

Sábado 03 de octubre, 9:00 a 12:00 hrs.

1. R puesto en práctica
2. Sintaxis y operaciones básicas
3. Manipulación de bases de datos

Instructor: Mauricio Bucca



Clase 04: Taller #2, Workflow

Jueves 08 de octubre, 18:00 a 20:00 hrs.

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

Instructor: Mauricio Bucca

Clase 05: Manipulación de bases de datos con tidyverse #1

Jueves 15 de octubre, 18:00 a 20:00 hrs.

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

Instructor: Mauricio Bucca

Clase 06: Manipulación de bases de datos con tidyverse #2

Jueves 22 de octubre, 18:00 a 20:00 hrs.

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

Instructor: Mauricio Bucca

Clase 07: Manipulación de bases de datos con tidyverse #3

Sábado 24 de octubre, 9:00 a 12:00 hrs.

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

Instructor: Mauricio Bucca



Clase 08: Taller #3, Manipulación de datos CASEN 2017 con herramientas tidyverse

Jueves 29 de octubre, 18:00 a 20:00 hrs.

1. Tidyverse puesto en práctica
2. Sintaxis y operaciones básicas
3. Manipulación de bases de datos

Instructor: Mauricio Bucca

Clase 09: Re-ordenación de bases de datos

Jueves 05 de noviembre, 18:00 a 20:00 hrs.

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

Instructor: Mauricio Bucca

Clase 10: Tratamiento de datos faltantes

Sábado 07 de noviembre, 9:00 a 12:00 hrs.

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

Instructor: Mauricio Bucca

Clase 11: Visualización de datos con ggplot2 #1

Jueves 19 de noviembre, 18:00 a 20:00 hrs.

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

Instructor: Mauricio Bucca

Clase 12: Visualización de datos con ggplot2 #2

Sábado 21 de noviembre, 9:00 a 12:00 hrs.

1. Personalización de gráficos
2. Gráficos por grupo



3. Ejemplos de gráficos avanzados
4. Exportación de figuras

Instructor: Mauricio Bucca

Clase 13: Automatización

Jueves 26 de noviembre, 18:00 a 20:00 hrs.

1. Funciones personalizadas
2. Herramientas de iteración

Instructor: Mauricio Bucca

Clase 14: Inferencia y test estadísticos básicos en R

Jueves 12 de noviembre, 18:00 a 20:00 hrs.

1. Error Estándar e Intervalo de confianza
2. Ilustración de la teoría vía simulación en R
3. Intervalo de confianza y prueba de hipótesis para una media y diferencia de medias
4. Intervalo de confianza y prueba de hipótesis para una proporción y diferencia de proporciones
5. Tablas de contingencia y test de independencia

Instructor: Mauricio Bucca

Clase 15: Reportes automatizados

Jueves 03 de diciembre, 18:00 a 20:00 hrs.

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

Instructor: Mauricio Bucca

Clase 16: Taller #4, Workflow avanzado

Jueves 10 de diciembre, 18:00 a 20:00 hrs.

1. Uso avanzado de Scripts
2. Construcción de un “workflow” efectivo, ordenado y automático



3. Exportación avanzada de resultados
4. Replicabilidad
5. Editores de texto

Instructor: Mauricio Bucca

Clase 17: Recapitulación

Sábado 12 de diciembre, 09:00 a 12:00 hrs.

1. Preguntas
2. Implementación en R de preguntas de estudiantes
3. Panorámica de otras herramientas útiles

Instructor: Mauricio Bucca

Cápsulas Formativas

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