INSTITUTO DE SOCIOLOGÍA
FACULTAD DE CIENCIAS SOCIALES

SOL4001
Procesamiento Avanzado de Bases de Datos con R

Mauricio Bucca (mebucca@uc.cl)

27 de agosto de 2020

CURSO	: **Procesamiento Avanzado de Bases de Datos**
NOMBRE INGLÉS	: Advanced Processing of Databases
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 Magíster, 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 secundarias, tanto en términos de manejo de variables como de consolidación de bases de datos (creación, traspaso de archivos, combinar archivos, cambiar de unidad de análisis). Además, los alumnos podrán tener la experiencia de trabajar con las más importantes bases de datos de encuestas disponibles de manera gratuita en y/o sobre Chile. Al final del curso se espera que los alumnos puedan enfrentar demandas de mediana a avanzada complejidad que impliquen el procesamiento de primera mano de distintas bases de datos disponibles en el país.

El desarrollo de los contenidos será en el programa R. Este es el software estadístico más utilizado en el mundo, ofrece un ambiente potente y flexible para el análisis de datos, contiene numerosas librerías para efectuar los análisis más recientes y novedoso, adicionalmente facilita los trabajos de visualización gráfica de los datos.

Será de responsabilidad de cada alumno descargar e instalar la última versión de los software R y RStudio. Descargar desde https://cran.dcc.uchile.cl/ y https://rstudio.com/products/rstudio/download/, respectivamente. Complementariamente, los estudiantes también podrán utilizar la R-Cloud, que se encuentra disponible online en https://rstudio.cloud/. Las instrucciones para acceder a la plataforma se revisará en la primera clase del curso.

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 (en vivo): Clases expositivas en directo online, via *streeming*, 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 (cápsulas formativas): clases expositivas y/o tutoriales disponibles a través de videos pre-grabados.

Horas de Ayudantía

El ayudante del curso mantendrá horario de consulta via Zoom los días lunes de cada semana entre las 18.00 y 19:00 hrs. Los estudiantes deberán agendar la reunión via email con al menos un día de anticipación.

Todas las **clases en vivo** serán grabadas y estarán disponibles por 14 días corridos, a través de un enlace en la plataforma web Classroom.

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

El acceso a la plataforma web del curso Classroom se cerrará 14 días después de la última actividad del curso, ya sea esta una clase o la entrega de un trabajo final. Se recomienda a los alumnos descargar periódicamente el material lectivo del curso para sus registros.

Se recuerda a los alumnos que las lecciones dictadas en la Universidad, y anotadas o recogidas de cualquier forma por aquellos alumnos a quienes van dirigidas, no podrán ser publicadas, total o parcialmente, sin autorización de los académicos autores de las mismas[1]. Las limitaciones a los usos que los alumnos pueden hacer del material de clases, incluido el material de video grabado durante las reuniones por plataforma Zoom y las cápsulas de video pre-grabadas, está considerada expresamente en el Reglamento de Propiedad Intelectual UC[2], el cual aplica tanto a los alumnos de Postgrado UC como a los alumnos de Educación Continua UC[3]. Se recuerda a todos los alumnos revisar los reglamentos que norman el funcionamiento de los miembros de la comunidad UC, los cuales están disponibles en la plataforma web del curso Classroom (Clase 00/Reglamento Alumnos).

V. EVALUACIÓN

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

---

[1] Ver el Art. 14 del Reglamento de Propiedad Intelectual UC, donde se explicita que se consideran "obras" protegidas por el derecho de autor, entre otras enumeradas "(...) Las conferencias, discursos, lecciones, memorias, comentarios y obras de la misma naturaleza tanto en la forma oral como en sus versiones escritas o grabadas;".

[2] Ver el Art. 19 del Reglamento de Propiedad Intelectual UC: "Las lecciones dictadas en la Universidad, y anotadas o recogidas de cualquier forma por aquellos alumnos a quienes van dirigidas, no podrán ser publicadas, total o parcialmente, sin autorización de los académicos autores de las mismas."

[3] Ver el Art. 51 del Reglamento del Alumno de Educación Continua UC: "El alumno deberá cumplir con los requerimientos del Programa en todo momento, desde su ingreso y durante toda su permanencia, y respetar las normas de honestidad académica y de convivencia vigentes en la Universidad. Para estos efectos, los alumnos de educación continua, se consideran parte de la comunidad universitaria y deberán adherir al Código de Honor de la Universidad."

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. La fecha de publicación y de entrega de cada una de las tareas está especificado 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

Examen final (40%)

El examen final pondera un **40% de la nota final del curso**, y se desarrolla en forma individual.

- Trabajo Final: Se asignará 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 post-grado. Otras posibles infracciones a la honestidad académica también serán derivadas a la subdirección donde se evaluarán posibles sanciones (ver Reglamento del Alumnos de Post-grado).

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 se sugerirá lecturas para reforzar y complementar lo revisado en clases.

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.,

Metodología de encuestas:

- Encuesta CASEN 2017: http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/casen_2017.php
- Encuesta Panel CASEN: http://observatorio.ministeriodesarrollosocial.gob.cl/enc_panel_basedatos.php
- Encuesta de protección social: https://www.previsionsocial.gob.cl/sps/biblioteca/encuesta-de-proteccion-social/bases-de-datos-eps/
- Encuesta Nacional de Empleo: https://www.ine.cl/estadisticas/laborales/ene

VIII. PROGRAMA DETALLADO DE CLASES

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

Sábado 26 de septiembre, 9:00 a 11:50 hrs.
Instructor: Mauricio Bucca

Contenidos:

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 02: Introducción a bases de datos

Jueves 01 de octubre de 2020, 18:00 a 20:50 hrs.
Instructor: Mauricio Bucca

Tarea #1:

Publicación: 01 de octubre de 2020, 23:55 hrs
Entrega: 08 de octubre de 2020, 23:55 hrs

Contenidos:

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

Clase 03: Taller #1, Primer acercamiento a base de datos en R. CASEN 2017

Sábado 03 de octubre, 9:00 a 11:50 hrs.
Instructor: Mauricio Bucca

Clase 04: Taller #2, Workflow

Jueves 08 de octubre de 2020, 18:00 a 20:50 hrs.
Instructor: Mauricio Bucca

Tarea #2:

Publicación: 08 de octubre de 2020, 23:55 hrs
Entrega: 15 de octubre de 2020, 23:55 hrs

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

Clase 05: Manipulación de bases de datos con tidyverse
Jueves 15 de octubre de 2020, 18:00 a 20:50 hrs.
Instructor: Mauricio Bucca

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

Clase 06: Manipulación de bases de datos con tidyverse
Jueves 22 de octubre de 2020, 18:00 a 20:50 hrs.
Instructor: Mauricio Bucca

Tarea #3:
Publicación: 22 de octubre de 2020, 23:55 hrs
Entrega: 29 de octubre de 2020, 23:55 hrs

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

Clase 07: Manipulación de bases de datos con tidyverse
Sábado 24 de octubre de 2020, 9:00 a 11:50 hrs.
Instructor: Mauricio Bucca

Contenidos:
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 un llave común

Clase 08: Taller #3, Manipulación de datos CASEN 2017 con herramientas tidyverse
Jueves 29 de octubre de 2020, 18:00 a 20:50 hrs.
Instructor: Mauricio Bucca

Clase 09: Re-ordenación de bases de datos
Jueves 05 de noviembre de 2020, 18:00 a 20:50 hrs.
Instructor: Mauricio Bucca

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

Clase 10: Tratamiento de datos faltantes

Sábado 07 de noviembre de 2020, 9:00 a 11:50 hrs.
Instructor: Mauricio Bucca

Tarea #4:

Publicación: 07 de noviembre de 2020, 23:55 hrs
Entrega: 14 de noviembre de 2020 de 2020, 23:55 hrs

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

Clase 11: Visualización de datos con ggplot2

Jueves 12 de noviembre de 2020, 18:00 a 20:50 hrs.
Instructor: Mauricio Bucca

Tarea #5:

Publicación: 12 de noviembre de 2020, 23:55 hrs
Entrega: 19 de noviembre de 2020, 23:55 hrs

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

Clase 12: Visualización de datos con ggplot2

Jueves 19 de noviembre de 2020, 18:00 a 20:50 hrs.
Instructor: Mauricio Bucca

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

Clase 13: Automatización

Sábado 21 de noviembre de 2020, 9:00 a 11:50 hrs.
Instructor: Mauricio Bucca

Contenidos:

1. Funciones personalizadas
2. Herramientas de iteración

Clase 14: Inferencia y test estadísticos básicos en R

Jueves 26 de noviembre de 2020, 18:00 a 20:50 hrs.
Instructor: Mauricio Bucca

Trabajo Final: xxxxxx

Publicación: 26 de noviembre de 2020, 23:55 hrs
Entrega: 17 de diciembre de 2020, 23:55 hrs

Contenidos:

1. Error Estándar e Intervalo de confianza.
2. Ilustración de la teoría via 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

Clase 15: Reportes automatizados

Jueves 03 de diciembre de 2020, 18:00 a 20:50 hrs.
Instructor: Mauricio Bucca

Contenidos:

1. RMarkdown y librería knirt
2. Escritura de reportes automarizados y replicables

Clase 16: Taller #4, Workflow avanzado

Jueves 10 de diciembre de 2020, 18:00 a 20:00 hrs.
Instructor: Mauricio Bucca

Contenidos:

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

Clase 17: Recapitulación

Sábado 12 de diciembre de 2020, 09:00 a 11:50 hrs.
Instructor: Mauricio Bucca

Contenidos:

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