

Gestión de datos epidemiológicos con tidyverse

Fundamentación

El análisis de datos inicia con el proceso de inspeccionar, depurar y transformar datos. En la práctica, un investigador gasta mucho tiempo en la preparación de los datos antes de realizar cualquier operación estadística.

No es habitual que los datos en bruto cumplan con todos los requisitos deseados, por lo que la depuración de datos (cleaning data) viene a dar solución como etapa de transformación para conseguir datos coherentes que se puedan analizar.

El lenguaje R es un potente entorno donde realizar todo este proceso y tidyverse un novedoso paquete que reúne poderosas herramientas orientadas a la gestión de datos.

Contribución esperada

Se espera que esta actividad proporcione un ambiente de trabajo y nuevas herramientas para la gestión de datos, su depuración y visualización.

Perfil del participante

Esta actividad está dirigida a aquellos profesionales y técnicos que trabajan con datos crudos como insumo de análisis, provengan estos de fuentes primarias o secundarias y en función de actividades de investigación o vigilancia epidemiológica.

Deberán poseer conocimientos básicos del lenguaje R.

Objetivos

Se espera que los participantes logren:

- Leer bases de datos de diferentes formatos.
- Construir procesos de depuración variable por variable y cruzada.
- Filtrar, clasificar, agrupar, ordenar y resumir datos de tablas leídas
- Organizar adecuadamente la información en tablas de datos técnicamente correctas y estadísticamente coherentes.
- Manipular formatos adecuados en variables de tiempo, lugar y persona.
- Crear nuevas variables calculadas a partir de las recolectadas originalmente.
- Generar visualizaciones estadísticas descriptivas de datos

Contenido

Unidad 1

Introducción: Conceptos básicos sobre ciencia de datos

Paquete tidyverse y su ambiente

Tuberías con magrittr

Unidad 2

Lectura de datos con readr
Datos ordenados con tidyr
Dataframes con tibble

Unidad 3

Manipulación de datos con dplyr
Trabajar fechas con lubridate
Trabajar cadenas de caracteres con stringr

Unidad 4

Visualización de datos con ggplot

Unidad 5

Programación funcional con purrr
Otros paquetes y funciones secundarias

Estrategias metodológicas y recursos didácticos

El docente se encargará del dictado de clases presenciales, con apoyo en dispositivos y ejercicios / tutoriales prácticos. También mantendrá un canal de comunicación para consultas por medio de un foro en aula virtual.

Los participantes realizarán:

- Actividades iniciales: presentaciones personales y situación laboral en relación a las herramientas informáticas propuestas.
- Actividades de desarrollo: lectura de material teórico/práctico, trabajos prácticos reforzando contenido, y elaboración de productos parciales.
- Actividades integradoras: realización de un informe/documento final integrador de todas las habilidades y conocimientos adquiridos durante el curso, encuesta de satisfacción sobre el curso.

Los alumnos contarán con los siguientes recursos: material de estudio seleccionado y presentado a través de la plataforma virtual ANLIS, espacio de consultas presencial en cada clase, foro de consultas para comunicarse con el docente entre clases, computadoras en aula informática del INE.

Descripción de la modalidad

Presencial

Bibliografía

Garrett Grolemund, Hadley Wickham – R for Data Science - O'Reilly - 2017

Chester Ismay, Albert Y. Kim - An Introduction to Statistical and Data Sciences via R - ModernDive - 2017

Evaluación de los aprendizajes

Evaluación de proceso

- Los alumnos deberán cumplir con la realización de las actividades planteadas a lo largo de la cursada.
- Se evaluarán las habilidades parciales que se vayan adquiriendo a través de las actividades propuestas para cada tema.

Evaluación de producto:

- El participante que haya completado y aprobado el total de los trabajos prácticos parciales, estará en condiciones de completar el curso con un trabajo práctico final que integrará todas las habilidades adquiridas a lo largo del curso.
- El trabajo práctico final será la elaboración de un script codificado en lenguaje R con el que depure, gestione y visualice datos de una base de datos epidemiológicos ficticios.

Instrumentos para la evaluación

Instrumentos para la evaluación del aprendizaje:

- Trabajos prácticos parciales
- Trabajo práctico final de integración de los conocimientos adquiridos con aplicación al uso laboral.

Instrumentos para la evaluación de la actividad:

- Encuesta de satisfacción realizada a cada alumno.

Requisitos de Asistencia y aprobación

Los participantes deberán:

- Cursar y participar en todas las unidades del curso, respetando las consignas propuestas.

- Realizar y aprobar los trabajos prácticos de cada proceso en los tiempos indicados.
- Aprobar el trabajo final de integración de los conocimientos adquiridos.

Duración (Hs.)

50 hs

Detalle sobre la duración

5 encuentros presenciales de 4 horas cada uno (20 hs)

Trabajos prácticos parciales (10 hs)

Trabajo práctico final será la elaboración de un informe o documento científico (20 hs)

Lugar:

-Plataforma educativa ANLIS

-Instituto Nacional de Epidemiología "Dr. Juan H. Jara"

Perfil instructor

Especialistas en el lenguaje R con conocimientos de epidemiología.

Docente

Ballejo Christian. Especialista en Sistemas, Servicio de Informática y Estadística, Departamento de Investigación Epidemiológica, INE "Dr. Juan H. Jara".

CUIL: 23-23313346-9