Maestría en Periodismo sobre Políticas Públicas

Visualización y Manejo de datos

Clases: viernes, 9-12 hrs.
Hrs. de oficina: martes, 11-13 hrs.

Sebastián Garrido de Sierra
sebastian.garrido@cide.edu

Introducción

Este curso está diseñado para fortalecer tus habilidades en dos áreas cada vez más importantes en el mercado laboral y que pocas veces se enseñan en un curso de licenciatura o posgrado: la **visualización** y el **manejo** de datos.

La **visualización de datos** es el área del conocimiento que combina principios de diseño, teoría del color y nuestras capacidades perceptivas para elaborar gráficas y mapas que transmiten de forma eficiente y efectiva los resultados sustantivos derivados del análisis de bases de datos.

En el contexto de esta clase, el **manejo de datos** incluye la construcción, limpieza y transformación[1] de bases de datos para su posterior uso. Dado que es inevitable que tengamos que dedicar una enorme cantidad de tiempo a estas tres tareas antes de poder analizar y/o visualizar nuestros datos, es crucial aprender principios y herramientas que nos permitan realizarlas de forma eficiente, ordenada y replicable/auditable.

Objetivos

A lo largo de este curso los alumnos aprenderán, entre otras cosas:

1. Cómo funcionan y se relacionan el sistema de la vista, nuestras percepciones y la memoria;

[1] En términos estrictos, el manejo de datos sólo incluye la transformación de los mismos. Sin embargo, en este curso aprenderás principios teóricos y habilidades prácticas para construir y limpiar bases de datos.

2. Principios teóricos básicos para visualizar datos de forma efectiva y eficiente;

3. Diversas técnicas y sistemas de visualización, así como cuándo es conveniente utilizar cada uno de ellos;

4. Cómo diseñar visualizaciones a través de un proceso estructurado, incorporando los principios de la percepción humana; y,

5. Cómo generar visualizaciones estáticas e interactiva utilizando diversos programas y plataformas web. La lista incluye, entre otros, *R*, Excel, LibreOffice, Tableau, CartoDB y Gephi.

6. Principios básicos para el diseño y construcción de bases de datos;

7. Cómo usar herramientas para diagnosticar si una base de datos tiene problemas y cómo corregirlos;

8. Qué es *R*, *RStudio* y los conocimientos básicos para utilizar estos programas;

9. Cómo manejar y transformar una base de datos en *RStudio* para, entre otras cosas:

    a. Crear nuevas variables construidas a partir de variables existentes;
    b. Renombrar variables;
    c. Seleccionar un subconjunto de renglones o columnas;
    d. Unir dos o más bases de datos;
    e. Calcular estadísticas descriptivas; y,
    f. Un largo etcétera...

Público objetivo

El curso está diseñado para personas con o sin experiencia en la manipulación de bases de datos y/o en la elaboración de visualizaciones. Los únicos prerrequisitos son:

1. Que los asistentes tengan conocimientos básicos de estadística;

2. Que sepan utilizar hojas de cálculo (Excel, LibreOffice, etc.);

3. Que estén dispuestos a aprender a programar en *R* y que hayan tomado este curso (http://bit.ly/1FORUxq) en línea gratuito **antes** de que comencemos a trabajar con *R* y *R Studio* en la sesión 8 (marzo 23).

4. Que hayan descargado e instalado R (https://cran.itam.mx) y RStudio (http://bit.ly/1Hllr0q), un entorno de desarrollo integrado que nos facilitará (dentro de lo que cabe) la vida al usar *R*.

Evaluación

Dada la naturaleza de este curso, la práctica cotidiana es clave. Por ello, tendrán al menos una tarea a la semana. En conjunto, estas tareas se traducirán en el **100%** de tu calificación semestral.

Sesiones

A lo largo del semestre tendremos 14 sesiones. Nos reuniremos cada viernes de 9 a 12 horas en el Teatro de Decisiones del Laboratorio Nacional de Políticas Públicas (LNPP), mismo que está ubicado en el tercer piso de la biblioteca. En caso de que el Teatro sea requerido para algún evento del LNPP, la clase será en alguno de los salones proporcionados por la Maestría de Periodismo sobre Políticas Públicas. Si esto ocurriera, se los notificaré con anticipación.

La mayoría de las 14 sesiones estarán divididas en dos partes. En la primera discutiremos aspectos teóricos del tema en cuestión y en la segunda los estudiantes aprenderán a usar diversas herramientas.

A continuación enlisto los temas que cubriremos en cada sesión. Mientras que las sesiones en **rojo** corresponden a clases en las que abordaremos temas relacionados con la **visualización de datos**, en las sesiones en **azul** cubriremos uno o más temas relacionados con el **manejo de datos** y *RStudio*. Al final de este documento encontrarás las fuentes bibliográficas de las cuales asignaré diversos capítulos.

Sesión 1 (ago. 25)

- Introducción al curso
- Introducción a la visualización de datos

Lecturas: ninguna. La vida aún es bella.

Sesión 2 (sep. 1)

- Entendiendo al ojo y el “cerebro visual”
- Reconociendo al recordar

Lecturas: Cairo, cap. 5, 6 y 7

Sesión 3 (sep. 8)

- ¿Qué?
- ¿Por qué?

Lecturas: Munzner, cap. 2 y 3

Sesión 4 (sep. 15)

- ¿Cómo?
- Organización gráfica I: tablas

Lecturas: Munzner, cap. 5.

Sesión 5 (sep. 22)

- Organización gráfica II: redes
- Organización gráfica III: mapas

Lecturas: ninguna. La vida otra vez es bella.

Sesión 6 (sep. 29)

- Principios de diseño
- Introducción a la segunda parte del curso

Lecturas: Tufte, cap. 2, 4 y 5.

Sesión 7 (oct. 6)

- Construcción y limpieza de bases de datos

Lecturas: Wickham, *Tidy Data*, pp. 1-5, url: http://bit.ly/2vMKMqp

Sesión 8 (oct. 13)

- Introducción a *R* y *RStudio*
- Tipos de datos

Lecturas: Phillips, sección 9.3

Sesión 9 (oct. 19)

- Tipos de estructuras de datos
- Cómo cargar base de datos en formatos .csv, .xlsx, .dat, .sav, etc.

Lecturas: Wickham y Grolemund, cap. 10 y 11 || Phillips, cap. 5, 6 y 8

Sesión 10 (oct. 26)

- Cómo “rebanar” bases de datos a la antigüita (R base)
- Cómo “rebanar”, transformar y analizar datos a la **`dplyr`**

Lecturas: Wickham y Grolemund, cap. 5 || Phillips, sección 10.4

Sesión 11 (nov. 10)

- Unión de bases de datos con **`dplyr`**
- *Tidyear* bases de datos con **`tidyr`**
- Unir y *tidyear*, o de cómo **`dplyr`** + **`tidyr`** = ❤

Lecturas: Wickham y Grolemund, caps. 12 y 13

Sesión 12 (nov. 17)

- Análisis de encuestas en *R*. Profesor invitado: Javier Márquez

Lecturas: Kreuter y Valliant, "A survey on survey statistics: What is done and can be done in Stata", url: http://bit.ly/2wr1Zmw.

Sesión 13 (nov. 24)

- Gráficas en R base
- Gráficas con **`ggplot2`** – I

Lecturas: Wickham y Grolemund, cap. 3

Sesión 14 (dic. 1)

- Gráficas con **`ggplot2`** – II
- Cómo reducir el caos: los proyectos de RStudio

Lecturas: Wickham y Grolemund, cap. 8 y 28

Bibliografía

Cairo, Alberto (2012) *The Functional Art: An Introduction to Information Graphics and Visualization*, New Riders, California, Estados Unidos.

Chang, Winston (2013) *R Graphics Cookbook*, O'Reilly Media, California, Estados Unidos.

Grolemund, Garrett (2014) *Hands-On Programming with R. Write Your Own Functions and Simulations*, O'Reilly Media, California, Estados Unidos.

Horton, Nicholas J., Randall Pruim y Daniel T. Kaplan (2015) *A Student's Guide to R*, Project MOSAIC.

Kreuter, Frauke y Richard Valliant (2007) "A survey on survey statistics: What is done and can be done in Stata", *The Stata Journal*, Núm. 1, pp. 1-21, url: http://bit.ly/2wr1Zmw.

Munzner, Tamara (2014) *Visualization Analysis and Design*, CRC Press, Nueva York, Estados Unidos.

Teetor, Paul (2011) *R Cookbook*, O'Reilly Media, California, Estados Unidos

Tufte, Edward R. (2001) *The Visual Display of Quantitative Information 2nd Edition*, Graphics Press, Connecticut, Estados Unidos.

Wickham, Hadley y Garrett Grolemund (2017) *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, O'Reilly Media, California, Estados Unidos.

Wickham, Hadley (2014) "Tidy Data", *Journal of Statistical Software*, Vol. 59, Issue 10, url: http://bit.ly/2vMKMqp .

Wilkinson, Leland (2005) *The Grammar of Graphics*, Springer, Canadá.