

Ejercicio 2: Análisis Visual y Exploratorio.

Introducción:

En este ejercicio, se debe identificar un conjunto de datos de interés y realizar un análisis inicial para comprender mejor la forma y estructura de los datos, investigar preguntas iniciales, y desarrollar una intuición e hipótesis preliminar. El resultado del trabajo debería tener la forma de un informe que consista principalmente en un conjunto de gráficos/imágenes subtituladas, donde se transmiten los conocimientos claves obtenidos durante su análisis.

Datos a utilizar:

En primera instancia, elija un área temática de su interés, y busque un conjunto de datos que le pueda brindar información sobre ese tema. En su defecto, puede considerar elegir algún conjunto de datos dentro de los que se especifican más abajo.

Sin embargo, si desea investigar un conjunto de datos diferente, puede hacerlo. Tenga en cuenta que la recopilación y preparación de datos puede ser un proceso tedioso.

Después de seleccionar el tema y el conjunto de datos, **debe escribir un conjunto inicial de preguntas** que le gustaría responder investigar.

Análisis visual exploratorio:

Se realizará un análisis exploratorio de su conjunto de datos utilizando R o alguna otra herramienta como Excel u Octave. Debe considerar dos fases diferentes de exploración:

1. Busque obtener una descripción general de la forma y estructura del conjunto de datos. ¿Qué variables contiene el conjunto de datos? ¿Cómo se distribuyen? ¿Hay problemas de calidad de datos? ¿Existen relaciones extrañas entre las variables? Asegúrese de realizar "prueba de cordura" (sanity check) para corroborar cualquier patrón que espere que contengan los datos.
2. **Investigue sus preguntas iniciales, así como cualquier pregunta nueva que surja durante la exploración.** Para cada pregunta, cree una primera versión de la visualización que pueda proporcionar una respuesta útil. Luego, refine/mejore esa visualización para desarrollar una mejor perspectiva, explorar observaciones inesperadas o verificar la cordura de sus suposiciones.

Entregable:

El entregable debe consistir principalmente de una secuencia de imágenes, (8 o más) que detallen sus observaciones más importantes. Estas pueden incluir sorpresas (por ejemplo, problemas de calidad de datos), así como respuestas a las preguntas de análisis. Cuando corresponda, incluya ["visualizaciones anotadas"](#)

para guiar la atención de los espectadores y proporcionar contexto a las imágenes. Proporcione detalles suficientes para que cualquiera pueda leer su informe y comprender lo que ha aprendido sin estar familiarizado con el conjunto de datos. Para ayudar a definir el alcance de esta tarea, vea [este informe de ejemplo](#) que analiza datos de películas.

Cada gráfica debe incluir títulos y anotaciones descriptivas que destaquen los que se muestran en esa vista. Por ejemplo, las anotaciones podrían tomar la forma de etiquetas de texto, colores diferenciales, círculos que contengan información importante, flechas, etc. También puede incluir un título para cada imagen, ¡pero breve! Las anotaciones pueden agregarse a mano, utilizando herramientas de edición de imágenes (Acrobat Reader, Paint u otro).

Las gráficas pueden estar realizadas en R, Octave/Matlab, MSExcel, LibreOffice Calc, etc. Puede realizarlas combinando varios paquetes y editores gráficos. En la medida de lo posible, guarde los códigos/archivos creados para generar las gráficas, así como una breve explicación de cómo fueron creadas, en especial cuando no son la salida directa de un único código.

El informe, que contiene las imágenes a mostrar, así como las explicaciones complementarias, es un PDF que se puede redactar en MSWord, LibreOffice, R Markdown, LaTeX o HTML.

Debe enviar un .zip con el informe en PDF, el archivo original (.docx, .odt, .rmd, .tex, etc.), los datos utilizados (.csv, .json, etc. o un link a ellos), los “programas” utilizados para la generación (.r, .m, .xls, .ods, etc.).

Fecha y procedimiento de entrega

Entrega a realizar antes del martes 1/8/23. Debe enviar un .zip con el informe en PDF, el archivo original (.docx, .odt, .rmd, .tex, etc.), los datos utilizados (.csv, .json, etc. o un link a ellos), los “programas” utilizados para la generación (.r, .m, .xls, .ods, etc.). Envíe un mail a eduardof@fing.edu.uy o a edufer111@gmail.com. También puede enviar un link de una dirección de WeTransfer o Dropbox al zip.

Potenciales fuentes de datos

Datos del Banco Mundial, 1960-2017

El Banco Mundial lleva un registro de del desarrollo humano con indicadores como cambio climático, economía, educación, medio ambiente, igualdad de género, salud, ciencia y tecnología, desde 1960. Se pueden explorar los datos originales por [indicadores](#) o por [países](#). También se pueden ver este resumen de [20 indicadores del Banco Mundial](#).

Daily Weather in the U.S., 2017

Este banco de datos contiene medidas del clima de U.S. durante 2017, provistas por la **NOAA Daily Global Historical Climatology Network**. Estos datos han sido

transformados: algunas estaciones metereológicas con medidas aisladas han sido eliminadas. Vea el archivo [weather.txt](#) con descripciones para cada columna.

Data: [weather.csv.gz](#) (gzipped CSV)

Otras posibles fuentes de datos

- **Instituto Nacional de Estadísticas:** <https://www.ine.gub.uy/>
- **INUMET:** <https://www.inumet.gub.uy/tiempo/estado-actual>
- **Catálogo de datos abiertos (¿?):** <https://catalogodatos.gub.uy/group/estadisticos>
- [data.seattle.gov](#) - City of Seattle Open Data
- [data.wa.gov](#) - State of Washington Open Data
- [nwdata.org](#) - Open Data & Civic Tech Resources for the Pacific Northwest
- [data.gov](#) - U.S. Government Open Datasets
- [U.S. Census Bureau](#) - Census Datasets
- [IPUMS.org](#) - Integrated Census & Survey Data from around the World
- [Federal Elections Commission](#) - Campaign Finance & Expenditures
- [Federal Aviation Administration](#) - FAA Data & Research
- [fivethirtyeight.com](#) - Data and Code behind the Stories and Interactives
- [Buzzfeed News](#) - Open-source data from BuzzFeed's newsroom
- [Kaggle Datasets](#) - Datasets for Kaggle contests
- [List of datasets useful for course projects](#) - curated by Mike Freeman